arXiv:1308.2955v1 [math.ST] 13 Aug 2013

Community Detection in Sparse Random Networks

Nicolas Verzelen[1] and Ery Arias-Castro[2]

We consider the problem of detecting a tight community in a sparse random network. This is formalized as testing for the existence of a dense random subgraph in a random graph. Under the null hypothesis, the graph is a realization of an Erdös-Rényi graph on N vertices and with connection probability p_0; under the alternative, there is an unknown subgraph on n vertices where the connection probability is $p_1 > p_0$. In (Arias-Castro and Verzelen, 2012), we focused on the asymptotically *dense* regime where p_0 is large enough that $\log(1 \vee (np_0)^{-1}) = o(\log(N/n))$. We consider here the asymptotically *sparse* regime where p_0 is small enough that $\log(N/n) = O(\log(1 \vee (np_0)^{-1}))$. As before, we derive information theoretic lower bounds, and also establish the performance of various tests. Compared to our previous work (Arias-Castro and Verzelen, 2012), the arguments for the lower bounds are based on the same technology, but are substantially more technical in the details; also, the methods we study are different: besides a variant of the scan statistic, we study other statistics such as the size of the largest connected component, the number of triangles, the eigengap of the adjacency matrix, etc. Our detection bounds are sharp, except in the Poisson regime where we were not able to fully characterize the constant arising in the bound.

Keywords: community detection, detecting a dense subgraph, minimax hypothesis testing, Erdös-Rényi random graph, scan statistic, planted clique problem, largest connected component.

1 Introduction

Community detection refers to the problem identifying communities in networks, e.g., circles of friends in social networks, or groups of genes in graphs of gene co-occurrences (Bickel and Chen, 2009; Girvan and Newman, 2002; Lancichinetti and Fortunato, 2009; Newman, 2006; Newman and Girvan, 2004; Reichardt and Bornholdt, 2006). Although fueled by the increasing importance of graph models and network structures in applications, and the emergence of large-scale social networks on the Internet, the topic is much older in the social sciences, and the algorithmic aspect is very closely related to graph partitioning, a longstanding area in computer science. We refer the reader to the comprehensive survey paper of Fortunato (2010) for more examples and references.

By community detection we mean, here, something slightly different. Indeed, instead of aiming at extracting the community (or communities) from within the network, we simply focus on deciding whether or not there is a community at all. Therefore, instead of considering a problem of graph partitioning, or clustering, we consider a problem of testing statistical hypotheses. We observe an undirected graph $\mathcal{G} = (\mathcal{E}, \mathcal{V})$ with $N := |\mathcal{V}|$ nodes. Without loss of generality, we take $\mathcal{V} = [N] := \{1, \dots, N\}$. The corresponding adjacency matrix is denoted $\mathbf{W} = (W_{i,j}) \in \{0,1\}^{N \times N}$, where $W_{i,j} = 1$ if, and only if, $(i,j) \in \mathcal{E}$, meaning there is an edge between nodes $i, j \in \mathcal{V}$. Note that $\mathbf{W}$ is symmetric, and we assume that $W_{ii} = 0$ for all i. Under the null hypothesis, the graph $\mathcal{G}$ is a realization of $\mathbb{G}(N, p_0)$, the Erdös-Rényi random graph on N nodes with probability of connection $p_0 \in (0,1)$; equivalently, the upper diagonal entries of $\mathbf{W}$ are independent and identically distributed with $\mathbb{P}(W_{i,j} = 1) = p_0$ for any $i \neq j$. Under the alternative, there is a subset of nodes indexed by $S \subset \mathcal{V}$ such that $\mathbb{P}(W_{i,j} = 1) = p_1$ for any $i, j \in S$ with $i \neq j$, while $\mathbb{P}(W_{i,j} = 1) = p_0$ for

[1]INRA, UMR 729 MISTEA, F-34060 Montpellier, FRANCE
[2]Department of Mathematics, University of California, San Diego, USA

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **13 AUG 2013**	2. REPORT TYPE	3. DATES COVERED **00-00-2013 to 00-00-2013**

4. TITLE AND SUBTITLE **Community Detection in Sparse Random Networks**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of California, San Diego,Department of Mathematics,9500 Gilman Dr,La Jolla,CA,92093**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S)
	11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**58**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

any other pair $i \neq j$. We assume that $p_1 > p_0$, implying that the connectivity is stronger between nodes in S, so that S is an assortative community. The subset S is not known, although in most of the paper we assume that its size $n := |S|$ is known. Let H_0 denote the null hypothesis, which consists of $\mathbb{G}(N, p_0)$ and is therefore simple. And let H_S denote the alternative where S is the anomalous subset of nodes. We are testing H_0 versus $H_1 := \bigcup_{|S|=n} H_S$. We consider an asymptotic setting where

$$N \to \infty, \quad n = n(N) \to \infty, \quad n/N \to 0, \quad n/\log N \to \infty, \tag{1}$$

meaning the subgraph is small, but not too small. Also, the probabilities of connection, $p_0 = p_0(N)$ and $p_1 = p_1(N)$, may change with N — in fact, they will tend to zero in most of the paper.

Despite its potential relevance to applications, this problem has received considerably less attention. We mention the work of Wang et al. (2008) who, in a somewhat different model, propose a test based on a statistic similar to the modularity of Newman and Girvan (2004); the test is evaluated via simulations. Sun and Nobel (2008) Rukhin and Priebe (2012) consider a test based on the maximum number of edges among the subgraphs induced by the neighborhoods of the vertices in the graph; they obtain the limiting distribution of this statistic in the same model we consider here, with p_0 and p_1 fixed, and n is a power of N, and in the process show that their test reduces to the test based on the maximum degree. Closer in spirit to our own work, Butucea and Ingster (2011) study this testing problem in the case where p_0 and p_1 are fixed. A dynamic setting is considered in (Heard et al., 2010; Mongiovı et al., 2013; Park et al., 2013) where the goal is to detect changes in the graph structure over time.

Our previous work. We recently considered this testing problem in (Arias-Castro and Verzelen, 2012), focusing on the *dense* regime where $\log(1 \vee (np_0)^{-1}) = o(\log(N/n))$ or equivalently $p_0 \geq n^{-1}(n/N)^{o(1)}$. (For $a, b \in \mathbb{R}$, $a \wedge b$ denotes the minimum of a and b.) We obtained information theoretic lower bounds, and we proposed and analyzed a number of methods, both when p_0 is known and when it is unknown. (None of the methods we considered require knowledge of p_1.) In particular, a combination of the total degree test based on

$$W := \sum_{1 \leq i < j \leq N} W_{i,j} \ , \tag{2}$$

and the scan test based on

$$W_n^* := \max_{|S|=n} W_S, \qquad W_S := \sum_{i,j \in S, i<j} W_{i,j} \ , \tag{3}$$

was found to be asymptotically minimax optimal when p_0 is known and when n is not too small, specifically $n/\log N \to \infty$. This extends the results that Butucea and Ingster (2011) obtained for p_0 and p_1 fixed (and p_0 known). In that same paper, we also proposed and studied a convex relaxation of the scan test, based on the largest n-sparse eigenvalue of $\mathbf{W}^2$, inspired by related work of Berthet and Rigollet (2012).

Also in (Arias-Castro and Verzelen, 2012), we separately considered the very special case where $p_1 = 1$, meaning that the anomalous subgraph S is a clique. This is the case that Sun and Nobel (2008) consider motivated by a data mining application. We confirmed the intuition that the clique test, which is based on the size of the largest clique, is asymptotically minimax optimal. We note that the special case where $p_0 = 1/2$ and $p_1 = 1$ is often called the ‘planted clique problem’ and has received substantial attention in recent times (Alon et al., 1998; Dekel et al., 2011; Feige and Ron, 2010), where the main goal has been on designing polynomial-time algorithms that can detect (or even extract) the clique with high confidence. No such algorithm is known when the clique is of

size $n = o(\sqrt{N})$, even though in this setting the clique test can detect cliques of size $n \geq c \log N$ if $c > 2/\log 2$ is fixed.

Continuing our work, in the present paper we focus on the *sparse* regime where

$$-\log(np_0) \geq c_0 \log(N/n) \text{ for some constant } c_0 > 0. \tag{4}$$

Obviously, (4) implies that $np_0 \leq 1$. We define

$$\lambda_0 = Np_0, \qquad \lambda_1 = np_1, \tag{5}$$

and note that λ_0 and λ_1 may vary with N.

This paper. Compared to our previous work (Arias-Castro and Verzelen, 2012), the derivation of the various lower bounds here rely on the same general approach. Let $\mathbb{G}(N, p_0; n, p_1)$ denote the random graph obtained by choosing S uniformly at random among subsets of nodes of size n, and then generating the graph under the alternative with S being the anomalous subset. When deriving a lower bound, we first reduce the composite alternative to a simple alternative, by testing $H_0 : \mathbb{G}(N, p_0)$ versus $\bar{H}_1 := \mathbb{G}(N, p_0; n, p_1)$. Let L denote the corresponding likelihood ratio, i.e., $L = \binom{N}{n}^{-1} \sum_{|S|=n} L_S$, where L_S is the likelihood ratio for testing H_0 versus H_S. Then these hypotheses merge in the asymptote if, and only if, $L \to 1$ in probability under H_0. A variant of the so-called 'truncated likelihood' method, introduced by Butucea and Ingster (2011), consists in proving that $\mathbb{E}_0(\tilde{L}) \to 1$ and $\mathbb{E}_0(\tilde{L}^2) \to 1$, where $\tilde{L}$ is a truncated likelihood of the form $\tilde{L} = \binom{N}{n}^{-1} \sum_{|S|=n} L_S \mathbb{1}_{\Gamma_S}$, where Γ_S is a carefully chosen event. An important difference with our previous work is the more delicate choice of Γ_S, which here relies more directly on properties of the graph under consideration. We mention that we use a variant to show that H_0 and $\bar{H}_1$ do *not* separate in the limit. This could be shown by proving that the two graph models $\mathbb{G}(N, p_0)$ and $\mathbb{G}(N, p_0; n, p_1)$ are contiguous. The 'small subgraph conditioning' method of Robinson and Wormald (1992, 1994) — see the more recent exposition in (Wormald, 1999) — was designed for that purpose. For example, this is the method that Mossel et al. (2012) use to compare a Erdös-Rényi graph with a stochastic block model[3] with two blocks of equal size. This method does not seems directly applicable in the situations that we consider here, in part because the second moment of the likelihood ratio L tends to infinity at the limit of detection.

Again compared to our previous work (Arias-Castro and Verzelen, 2012), the methods we propose and study here are different. Although the total degree test (2) remains a contender, scanning over subsets of size exactly n as in (3) does not seem to be optimal anymore, all the more so when p_0 is small. Instead, we scan over subsets of a wider range of sizes, using

$$W_n^{\ddagger} = \sup_{k=n/u_N}^{n} \frac{W_k^*}{k} \,, \tag{6}$$

where $u_N = \log\log(N/n)$. We call this the broad scan test.

Our results. In analogy with our previous results in (Arias-Castro and Verzelen, 2012), we find that a combination of the total degree test (2) and the broad scan test based on (6) is asymptotically optimal when $\lambda_0 \to \infty$, in the following sense. Reparameterize $\lambda_0 = (N/n)^{\alpha}$ with $0 < \alpha \leq 1$ — the case $\lambda_0 \geq N/n$ being settled in (Arias-Castro and Verzelen, 2012) — and consider $n = N^{\kappa}$ with $0 < \kappa < 1$. Then, if $\kappa > \frac{1+\alpha}{2+\alpha}$, the total degree test is asymptotically powerful when $\lambda_1 \gg \frac{N^{(1+\alpha)/2}}{n^{1+\alpha}}$ and the two hypotheses merge asymptotically when $\lambda_1 \ll \frac{N^{(1+\alpha)/2}}{n^{1+\alpha}}$. When $\kappa < \frac{1+\alpha}{2+\alpha}$, that is for

[3]This is a popular model of a network with communities, also known as the planted partition model. In this model, the nodes belong to blocks: nodes in the same block connect with some probability p_{in}, while nodes in different blocks connect with probability p_{out}.

Table 1: Detection boundary and near-optimal algorithms in the regime $\lambda_0 = (N/n)^\alpha$ with $0 < \alpha < 1$ and $n = N^\kappa$ with $0 < \kappa < 1$. Undetectable means that that the hypotheses merge asymptotically, while detectable means that there exists an asymptotically powerful test. Here, $a \prec b$ (resp. $a \succ b$), means that there exists a positive constant C such that $a \le Cb$ (resp. $a \ge Cb$).

κ	$\kappa < \frac{1+\alpha}{2+\alpha}$	$\kappa > \frac{1+\alpha}{2+\alpha}$
Undetectable	$\lambda_1 \prec (1-\alpha)^{-1}$; Exact Eq. in (54)	$\lambda_1 \ll \frac{N^{(1+\alpha)/2}}{n^{1+\alpha}}$
Detectable	$\lambda_1 \succ (1-\alpha)^{-1}$; Exact Eq. in (13)	$\lambda_1 \gg \frac{N^{(1+\alpha)/2}}{n^{1+\alpha}}$
Optimal test	Broad Scan test	Total Degree test

Table 2: Detection boundary and near-optimal algorithms in the regimes $\lambda_0 \to \infty$ and $\lambda_0 \to 0$ and $n = N^\kappa$ with $0 < \kappa < 1/2$. For $1/2 < \kappa < 1$, the detection boundary accurs at $\lambda_1 \asymp N^{1/2}/n^2$ and is achieved by the total degree test.

	Undetectable	Detectable	Optimal test
$1 \ll \lambda_0 \ll \left(\frac{N}{n}\right)^{o(1)}$	$\limsup \lambda_1 < 1$	$\liminf \lambda_1 > 1$	Broad Scan test
$\frac{1}{N^{o(1)}} \le \lambda_0 = o(1)$	$\limsup \frac{\log(\lambda_1^{-1})}{\log(\lambda_0^{-1})} > \kappa$	$\liminf \frac{\log(\lambda_1^{-1})}{\log(\lambda_0^{-1})} < \kappa$	Largest CC test

smaller n, there exists a sequence of increasing functions ψ_n (defined in Theorem 1) such that the broad scan test is asymptotically powerful when $\liminf(1-\alpha)\psi_n(\lambda_1) > 1$ and the hypotheses merge asymptotically when $\limsup(1-\alpha)\psi_n(\lambda_1) < 1$. See Table 1 and the first line of Table 2 for a visual summary.

When $N^{-o(1)} \le \lambda_0 \le (N/n)^{o(1)}$ and $n = N^\kappa$ with $1/2 < \kappa < 1$, the total degree test is optimal, in the sense that it is asymptotically powerful for $\lambda_1^2/\lambda_0 \gg n^2/N$, while the hypotheses merge asymptotically — meaning all tests are asymptotically powerless — for $\lambda_1^2/\lambda_0 \ll n^2/N$. This is why we assume in the remainder of this discussion that $n = N^\kappa$ with $0 < \kappa < 1/2$.

The Poissonian regime where λ_0 and λ_1 are considered as fixed is depicted on Figure 1. When $\lambda_1 > 1$, the broad scan test is asymptotically powerful. When $\lambda_0 > e$ and $\lambda_1 < 1$, no test is able to fully separate the hypotheses. In fact, when λ_0 is bounded from above, and λ_1 is bounded from below away from 0, then the test based on the number of triangles has some nontrivial power, implying that the two hypotheses do not completely merge in this case. The case where $\lambda_0 < e$ is not completely settled. No test is able to fully separate the hypotheses if $\lambda_1 < \sqrt{\lambda_0/e}$. The largest connected component test is optimal up to a constant when $\lambda_0 < 1$ and a test based on counting subtrees of a certain size bridges the gap in constants for $1 \ge \lambda_0 < e$, but not completely. When λ_0 is bounded from above and $\lambda_1 = o(1)$, the two hypotheses merge asymptotically. Finally, when $\lambda_0 \to 0$, the largest connected component test is asymptotically optimal (See Table 2).

We also discuss tests that can be computed in polynomial time. Besides the total degree test, the test based on the largest connected component and the number of triangle test, already mentioned, we discuss the maximum degree test, a test based on counting simple cycles of given (small) length, and also some spectral methods. We find that, in the regime where $\lambda_0 \to \infty$, no test seems to come close to the broad scan test.

Content. The remaining of the paper is organized as follow. In Section 2 we introduce some notation and some concepts in probability and statistics, including concepts related to hypothesis

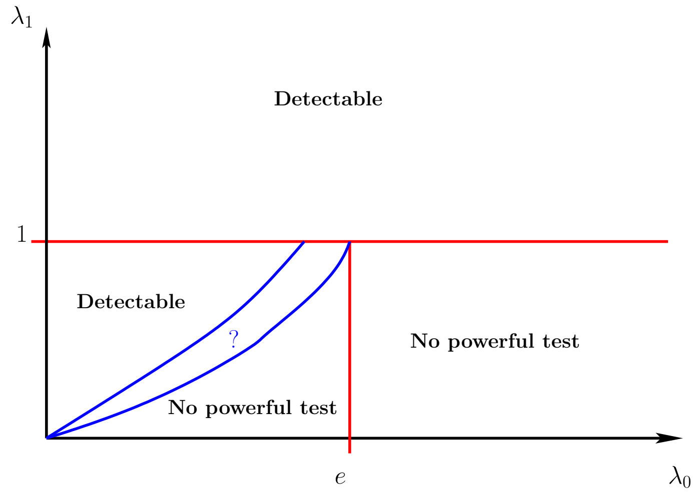

Figure 1: Detection diagram in the poissonian asymptotic where λ_0 and λ_1 are fixed and $n = N^\kappa$ with $0 < \kappa < 1/2$. "No powerful test" means that no test is able to fully separate the hypotheses.

testing and some basic results on the binomial distribution. in Section 3 we study some tests that are near-optimal in different regimes. In Section 4 we state and prove information theoretic lower bounds on the difficulty of the detection problem. In Section 5 we study various tests that run in polynomial-time. In Section 6 we discuss the situations where p_0 and/or n are unknown, as well as open problems. Section 7 contains some proofs and technical derivations.

2 Preliminaries

In this section, we first define some general assumptions and some notation, although more notation will be introduced as needed. We then define and discuss some basic concepts on hypothesis testing and decision theory, and in particular describe our strategy for obtaining information theoretic lower bounds. Finally, we list some general results that will be used multiple times throughout the paper.

2.1 Assumptions and notation

We recall that $N \to \infty$ and the other parameters such as n, p_0, p_1 may change with N, and this dependency is left implicit. We assume that p_0 is bounded away from 1, which is the most interesting case by far, and that $N^2 p_0 \to \infty$, the latter implying that the number of edges in the network (under the null) is not bounded. Similarly, we assume that $n^2 p_1 \to \infty$, for otherwise there is a non-vanishing chance that the community does not contain any edges. Unless stated otherwise, we assume that n and p_0 are both known.

Define

$$\alpha = \frac{\log \lambda_0}{\log(N/n)} \ , \tag{7}$$

in such a way that $p_0 = \frac{\lambda_0}{N}$ with $\lambda_0 = \left(\frac{N}{n}\right)^\alpha$. The dense regime considered in (Arias-Castro and Verzelen, 2012) corresponds to $\liminf \alpha \geq 1$. Here we focus on the sparse regime where $\limsup \alpha < 1$. The case where $\alpha \to 0$ includes the Poisson regime where λ_0 is constant.

Recall that $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is the (undirected, unweighted) graph that we observe, and for $S \subset \mathcal{V}$, let $\mathcal{G}_S$ denote the subgraph induced by S in $\mathcal{G}$.

We use standard notation such as $a_N \sim b_N$ when $a_N/b_N \to 1$; $a_N = o(b_N)$ when $a_N/b_N \to 0$; $a_N = O(b_N)$ when a_N/b_N is bounded; $a_N \asymp b_N$ when $a_N = O(b_N)$ and $b_N = O(a_N)$; $a_N \prec b_N$ when there exists a positive constant C such that $a_N \le Cb_N$ and $a_N \succ b_N$ when there exists a positive constant C such that $a_N \ge Cb_N$. We extend this notation to random variables. For example, if A_N and B_N are random variables, then $A_N \sim B_N$ if $A_N/B_N \to 1$ in probability.

For $x \in \mathbb{R}$, define $x_+ = x \vee 0$ and $x_- = (-x) \vee 0$, which are the positive and negative parts of x. For an integer n, let

$$n^{(2)} = \binom{n}{2} = \frac{n(n-1)}{2} \ . \tag{8}$$

Because of its importance in describing the tails of the binomial distribution, the following function — which is the relative entropy or Kullback-Leibler divergence of Bern(q) to Bern(p) — will appear in our results:

$$H_p(q) = q \log\left(\frac{q}{p}\right) + (1-q)\log\left(\frac{1-q}{1-p}\right), \quad p, q \in (0,1). \tag{9}$$

We let $H(q)$ denote $H_{p_0}(q)$.

2.2 Hypothesis testing

We start with some concepts related to hypothesis testing. We refer the reader to (Lehmann and Romano, 2005) for a thorough introduction to the subject. A test ϕ is a function that takes $\mathbf{W}$ as input and returns $\phi = 1$ to claim there is a community in the network, and $\phi = 0$ otherwise. The (worst-case) risk of a test ϕ is defined as

$$\gamma_N(\phi) = \mathbb{P}_0(\phi = 1) + \max_{|S|=n} \mathbb{P}_S(\phi = 0) \ ,$$

where $\mathbb{P}_0$ is the distribution under the null H_0 and $\mathbb{P}_S$ is the distribution under H_S, the alternative where S is anomalous. We say that a sequence of tests (ϕ_N) for a sequence of problems $(\mathbf{W}_N)$ is asymptotically powerful (resp. powerless) if $\gamma_N(\phi_N) \to 0$ (resp. $\to 1$). We will often speak of a test being powerful or powerless when in fact referring to a sequence of tests and its asymptotic power properties. Then, practically speaking, a test is asymptotically powerless if it does not perform substantially better than any method that ignores the adjacency matrix $\mathbf{W}$, i.e., guessing. We say that the hypotheses merge asymptotically if

$$\gamma_N^* := \inf_\phi \gamma_N(\phi) \to 1 \ ,$$

and that the hypotheses separate completely asymptotically if $\gamma_N^* \to 0$, which is equivalent to saying that there exists a sequence of asymptotically powerful tests. Note that if $\liminf \gamma_N^* > 0$, no sequence of tests is asymptotically powerful, which includes the special case where the two hypotheses are contiguous.

2.3 Some general results

Remember the definition of the entropy function in (9). The following is a simple concentration inequality for the binomial distribution.

Lemma 1 (Chernoff's bound)**.** *For any positive integer n, any $q, p \in (0,1)$, we have*

$$\mathbb{P}\left(Bin(n,p) \ge qn\right) \le \exp\left(-nH_p(q)\right). \tag{10}$$

Here are some asymptotics for the entropy function.

Lemma 2. *Define* $h(x) = x \log x - x + 1$. *For* $0 < p \leq q < 1$, *we have*

$$0 \leq H_p(q) - p\, h(q/p) \leq O(\frac{q^2}{1-q}) \ .$$

The following are standard bounds on the binomial coefficients. Recall that $e = \exp(1)$.

Lemma 3. *For any integers* $1 \leq k \leq n$,

$$\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \left(\frac{en}{k}\right)^k \ . \tag{11}$$

Let $\text{Hyp}(N, m, n)$ denotes the hypergeometric distribution counting the number of red balls in n draws from an urn containing m red balls out of N.

Lemma 4. $\text{Hyp}(N, m, n)$ *is stochastically smaller than* $\text{Bin}(n, \rho)$, *where* $\rho := \frac{m}{N-m}$.

3 Some near-optimal tests

In this section we consider several tests and establish their performances. We start by recalling the result we obtained for the total degree test, based on (2), in our previous work (Arias-Castro and Verzelen, 2012). Recalling the definition of λ_0 and λ_1 in (5), define

$$\zeta := \frac{(p_1 - p_0)^2}{p_0} \frac{n^4}{N^2} = \frac{(\lambda_1 - \lambda_0 n/N)^2}{\lambda_0} \frac{n^2}{N} \ . \tag{12}$$

Proposition 1 (Total degree test). *The total degree test is asymptotically powerful if* $\zeta \to \infty$, *and asymptotically powerless if* $\zeta \to 0$.

In view of Proposition 1, the setting becomes truly interesting when $\zeta \to 0$, which ensures that the naive total degree test is indeed powerless.

3.1 The broad scan

In the denser regimes that we considered in (Arias-Castro and Verzelen, 2012), the (standard) scan test based on W_n^* defined in (3) played a major role. In the sparser regimes we consider here, the broad scan test based on $W_n^\ddagger$ defined in (6) has more power. Assume that $\liminf \lambda_1 > 1$, so that $\mathcal{G}_S$ is supercritical under H_S. Then it is preferable to scan over the largest connected component in $\mathcal{G}_S$ rather than scan $\mathcal{G}_S$ itself.

Lemma 5. *For any* $\lambda > 1$, *let* η_λ *denote the smallest solution of the equation* $\eta = \exp(\lambda(\eta - 1))$. *Let* $\mathcal{C}_m$ *denote a largest connected component in* $\mathbb{G}(m, \lambda/m)$ *and assume that* $\lambda > 1$ *is fixed. Then, in probability,* $|\mathcal{C}_m| \sim (1 - \eta_\lambda)m$ *and* $W_{\mathcal{C}_m} \sim \frac{\lambda}{2}(1 - \eta_\lambda^2)m$.

Proof. The bounds on the number of vertices in the giant component is well-known (Van der Hofstad, 2012, Th. 4.8), while the lower bound on the number of edges comes from (Pittel and Wormald, 2005, Note 5). $\square$

By Lemma 5, most of the edges of $\mathcal{G}_S$ lie in its giant component, which is of size roughly $(1-\eta_{\lambda_1})n$. This informally explains why a test based on $W^*_{n(1-\eta_{\lambda_1})}$ is more promising that the standard scan test based on W^*_n.

In the details, the exact dependency of the optimal subset size to scan over seems rather intricate. This is why in $W^\ddagger_n$ we scan over subsets of size $n/u_N \le k \le n$. (Recall that $u_N = \log\log(N/n)$, although the exact form of u_N is not important.) For any subset $S \subset \mathcal{V}$, let

$$W^*_{k,S} = \max_{T\subset S, |T|=k} W_T \ .$$

Note that $W^*_{k,\mathcal{V}} = W^*_k$ defined in (3). Recall the definition of the exponent α in (7).

Theorem 1 (Broad scan test)**.** *The scan test based on $W^\ddagger_n$ is asymptotically powerful if either*

$$\limsup \alpha \le 1 \quad \text{and} \quad \liminf\ (1-\alpha) \sup_{k=n/u_N}^{n} \frac{\mathbb{E}_S[W^*_{k,S}]}{k} > 1 \ ; \tag{13}$$

or

$$\alpha \to 0 \quad \text{and} \quad \liminf \lambda_1 > 1 \ . \tag{14}$$

We shall prove in the next section that the power of the broad scan test is essentially optimal: if either $\limsup \alpha < 1$ and $\limsup\ (1-\alpha)\sup_{k=n/u_N}^{n} \mathbb{E}_S[W^*_{k,S}]/k < 1$, or $\alpha \to 0$ and $\limsup \lambda_1 < 1$, then no test is asymptotically powerful (at least when $n^2 = o(N)$, so that the total degree test is powerless). Regarding (13), we could not get a closed-form expression of this supremum. Nevertheless, we show in the proof that

$$\liminf \sup_{k=n/u_N}^{n} \frac{\mathbb{E}_S[W^*_{k,S}]}{k} \ge \liminf\ \frac{\lambda_1}{2}(1+\eta_{\lambda_1}) \ , \tag{15}$$

where η_λ is defined in Lemma 5. Moreover, we show in Section 7 the following upper bound.

Lemma 6.

$$\liminf \sup_{k=n/u_N}^{n} \frac{\mathbb{E}_S[W^*_{k,S}]}{k} \le \liminf\ \frac{\lambda_1}{2} + 1 + \sqrt{1+\lambda_1} \ . \tag{16}$$

Hence, assuming α and λ_1 are fixed and positive , the broad scan test is asymptotically powerful when $(1-\alpha)\frac{\lambda_1}{2}(1+\eta_{\lambda_1}) > 1$. In contrast, the scan test was proved to be asymptotically powerful when $(1-\alpha)\frac{\lambda_1}{2} > 1$ (Arias-Castro and Verzelen, 2012, Prop. 3), so that we have improved the bound by a factor larger than $1+\eta_{\lambda_1}$ and smaller than $1+2\lambda_1^{-1}(1+\sqrt{1+\lambda_1})$. When α converges to one, it was proved in (Arias-Castro and Verzelen, 2012) that the minimax detection boundary corresponds to $(1-\alpha)\lambda_1/2 \sim 1$ (at least when $n^2 = o(N)$). Thus, for α going to one, both the broad scan test and the scan test have comparable power and are essentially optimal. In the dense case, the broad scan test and the scan test have also comparable powers as shown by the next result which is the counterpart of (Arias-Castro and Verzelen, 2012, Prop. 3).

Proposition 2. *Assume that p_0 is bounded away from one. The broad scan test is powerful if*

$$\liminf \frac{nH(p_1)}{2\log(N/n)} > 1 \ .$$

The proof is essentially the same as the corresponding result for the scan test itself. See (Arias-Castro and Verzelen, 2012).

Proof of Theorem 1. First, we control $W_n^\ddagger$ under the null hypothesis. For any positive constant $c_0 > 0$, we shall prove that

$$\mathbb{P}_0\left[(1-\alpha)W_n^\ddagger \geq 1 + c_0\right] = o(1)\ . \tag{17}$$

Consider any integer $k \in [n/u_N, n]$, and let $q_k = 2(1+c_0)/[(k-1)(1-\alpha)]$. Recall that $k^{(2)} = k(k-1)/2$. Applying a union bound and Chernoff's bound (Lemma 1), we derive that

$$\begin{aligned} \mathbb{P}_0\left[W_k^* \geq \frac{1+c_0}{1-\alpha}k\right] &\leq \binom{N}{k}\exp\left[-k^{(2)}H(q_k)\right] \\ &\leq \exp\left[k\Big\{\log(eN/k) - \frac{k-1}{2}H(q_k)\Big\}\right]\ . \end{aligned}$$

We apply Lemma 2 knowing that $q_k/p_0 \to \infty$, and use the definition of α in (7), to control the entropy as follows

$$\begin{aligned} \frac{k-1}{2}H(q_k) &\sim \frac{k-1}{2}q_k \log\left[\frac{q_k}{p_0}\right] \\ &= \frac{1+c_0}{1-\alpha}\big[\log(N/n) - \log\lambda_0 + O(\log u_N)\big] \\ &\sim (1+c_0)\log(N/n)\ , \end{aligned}$$

since $\log(u_N) = o(\log(N/n))$. Consequently,

$$\mathbb{P}_0\left[W_k^* \geq \frac{1+c_0}{1-\alpha}k\right] \leq \exp\left[-kc_0\log(N/n)(1+o(1))\right]\ ,$$

where the $o(1)$ is uniform with respect to k. Applying a union bound, we conclude that

$$\mathbb{P}_0\left[(1-\alpha)W_n^\ddagger \geq 1+c_0\right] \leq \sum_{k=n/u_N}^{n} \exp\left[-kc_0\log(N/n)(1+o(1))\right] = o(1)\ .$$

We now lower bound $W_n^\ddagger$ under the alternative hypothesis. First, assume that (13) holds, so that there exists a positive constant c and a sequence of integer $k_n \geq n/u_N$ such that $\mathbb{E}_S[W_{k_n,S}^*] \geq k_n(1+c)/(1-\alpha)$ eventually. In particular, $\mathbb{E}_S[W_{k_n,S}^*] \to \infty$. We then use (19) in the following concentration result for $W_{k,S}^*$.

Lemma 7. *For any integer $0 \leq k \leq n$, we have the following deviation inequalities*

$$\mathbb{P}_S\left[W_{k,S}^* \geq \mathbb{E}_S[W_{k,S}^*] + t\right] \leq \exp\left[-\frac{\log(2)}{4}\left\{t \wedge \frac{t^2}{8E_S[W_{k,S}^*]}\right\}\right]\ , \quad \forall t > 8\left(1 \vee \sqrt{\mathbb{E}_S[W_{k,S}^*]}\right)\ ; \tag{18}$$

$$\mathbb{P}_S\left[W_{k,S}^* \leq \mathbb{E}_S[W_{k,S}^*] - t\right] \leq \exp\left[-\log(2)\frac{t^2}{8E_S[W_{k,S}^*]}\right]\ , \quad \forall t > 4\sqrt{\mathbb{E}_S[W_{k,S}^*]}\ . \tag{19}$$

It follows for Lemma 7 that, with probability going to one under $\mathbb{P}_S$,

$$W_n^\ddagger \geq \frac{W_{k_n}^*}{k_n} \geq \frac{W_{k_n,S}^*}{k_n} \geq \frac{1+c/2}{1-\alpha}\ .$$

Taking $c_0 = c/4$ in (17) allows us to conclude that the test based on $W_n^\ddagger$ with threshold $\frac{1+c/2}{1-\alpha}$ is asymptotically powerful.

Now, assume that (14) holds. Because $W_n^\ddagger$ is stochastically increasing in λ_1 under $\mathbb{P}_S$, we may assume that $\lambda_1 > 1$ is fixed. We use a different strategy which amounts to scanning the largest connected component of $\mathcal{G}_S$. Let $\mathcal{C}_{\max}^S$ be a largest connected component of $\mathcal{G}_S$.

For a small $c > 0$ to be chosen later, assume that $(1-c)n(1-\eta_{\lambda_1}) \leq |\mathcal{C}_{\max}^S| \leq (1+c)n(1-\eta_{\lambda_1})$ and $W_{\mathcal{C}_{\max}^S} \geq (1-c)\frac{n\lambda_1}{2}(1-\eta_{\lambda_1}^2)$, which happens with high probability under $\mathbb{P}_S$ by Lemma 5. Note that, because $\lambda_1 > 1$, we have $\eta_{\lambda_1} < 1$, and therefore $|\mathcal{C}_{\max}^S| \asymp n$. Consequently, when computing $W_n^\ddagger$ we scan $\mathcal{C}_{\max}^S$, implying that

$$W_n^\ddagger \geq \frac{W_{\mathcal{C}_{\max}^S}}{|\mathcal{C}_{\max}^S|} \geq \frac{(1-c)\frac{\lambda_1}{2}(1-\eta_{\lambda_1}^2)n}{(1+c)(1-\eta_{\lambda_1})n} \geq \frac{1-c}{1+c}\,\frac{\lambda_1}{2}(1+\eta_{\lambda_1}).$$

Since c above may be taken as small as we wish, and in view of (17), it suffices to show that $\lambda_1(1+\eta_{\lambda_1}) > 2$. Since η_λ converges to one when λ goes to one, we have $\lim_{\lambda\to 1}\lambda(1+\eta_\lambda) = 2$. Consequently, it suffices to show that the function $f : \lambda \mapsto \lambda(1+\eta_\lambda)$ is increasing on $(1,\infty)$. By definition of η_λ, we have $\eta_\lambda < 1/\lambda$ (since $e^{-\lambda} < 1/\lambda$) and $\eta'(\lambda) = \eta_\lambda(\eta_\lambda - 1)/(1-\lambda\eta_\lambda)$. Consequently, $f'(\lambda) = 2 + \frac{\eta_\lambda - 1}{1-\lambda\eta_\lambda}$. Hence, $f'(\lambda)$ is positive if $\eta_\lambda < (2\lambda-1)^{-1} := a_\lambda$. Recall that η_λ is the smallest solution of the equation $x = \exp[\lambda(x-1)]$, the largest solution being $x = 1$. Furthermore, we have $x \geq \exp[\lambda(x-1)]$ for any $x \in [\eta_\lambda, 1]$. To conclude, it suffices to prove $a_\lambda > e^{\lambda(a_\lambda - 1)}$. This last bound is equivalent to

$$\lambda - \frac{1}{2} - \frac{1}{2(2\lambda-1)} - \log(2\lambda-1) > 0\ .$$

The function on the LHS is null for $\lambda = 1$. Furthermore, its derivative $\frac{4(\lambda-1)^2}{(2\lambda-1)^2}$ is positive for $\lambda > 1$, which allows us conclude. $\square$

Proof of Lemma 7. The proof is based on moment bounds for functions of independent random variables due to Boucheron et al. (2005) that generalize the Efron-Stein inequality.

Recall that $\mathcal{G}_S = (S, \mathcal{E}_S)$ is the subgraph induced by S. Fix some integer $k \in [0, n]$. For any $(i,j) \in \mathcal{E}_S$, define the graph $\mathcal{G}_S^{(i,j)}$ by removing (i,j) from the edge set of $\mathcal{G}_S$. Let $W_T^{(i,j)}$ be defined as W_T^* but computed on $\mathcal{G}_S^{(i,j)}$, and then let $W_{k,S}^{*(i,j)} = \max_{T\subset S, |T|=k} W_T^{(i,j)}$. Observe that $0 \leq W_{k,S}^* - W_{k,S}^{*(i,j)} \leq 1$ and that $W_{k,S}^{*(i,j)}$ is a measurable function of $\mathcal{E}_S^{(i,j)}$, the edges set of $\mathcal{G}_S^{(i,j)}$. Let $T^* \subset S$ be a subset of size k such that $W_{k,S}^* = W_{T^*}$. Then, we have

$$\sum_{(i,j)\in\mathcal{E}_S}\left(W_{k,S}^* - W_{k,S}^{*(i,j)}\right) \leq \sum_{(i,j)\in\mathcal{E}_S}\left(W_{T^*} - W_{T^*}^{(i,j)}\right) = W_{T^*} = W_{k,S}^*\ ,$$

where the first equality comes from the fact that $W_{T^*} - W_{T^*}^{(i,j)} = \mathbb{1}_{\{(i,j)\in\mathcal{E}_T\}}$.

Applying (Boucheron et al., 2005, Cor. 1), we derive that, for any real $q \geq 2$,

$$\begin{aligned}\left[\mathbb{E}_S\left\{\left(W_{k,S}^* - \mathbb{E}_S[W_{k,S}^*]\right)_+^q\right\}\right]^{1/q} &\leq \sqrt{2q\,\mathbb{E}_S[W_{k,S}^*]} + q\ ;\\ \left[\mathbb{E}_S\left\{\left(W_{k,S}^* - \mathbb{E}_S[W_{k,S}^*]\right)_-^q\right\}\right]^{1/q} &\leq \sqrt{2q\,\mathbb{E}_S[W_{k,S}^*]}\ .\end{aligned}$$

Take some $t > 8(1 \vee \sqrt{\mathbb{E}_S[W_{k,S}^*]})$. For any $q \geq 2$, we have by Markov's inequality

$$\mathbb{P}_S\left[W_{k,S}^* \geq \mathbb{E}_S[W_{k,S}^*] + t\right] \leq \left(\frac{\sqrt{2q\,\mathbb{E}_S[W_{k,S}^*]} + q}{t}\right)^q\ .$$

The choice $q = \frac{t}{4} \wedge \frac{t^2}{32\,\mathbb{E}_S[W^*_{k,S}]}$ is larger than 2 and leads to (18). Similarly, if take some $t > 4\sqrt{\mathbb{E}_S[W^*_{k,S}]}$, and apply Markov's inequality, we get

$$\mathbb{P}_S\left[W^*_{k,S} \leq \mathbb{E}_S[W^*_{k,S}] - t\right] \leq \left(\frac{\sqrt{2q\,\mathbb{E}_S[W^*_{k,S}]}}{t}\right)^q .$$

The choice $q = \frac{t^2}{8\,\mathbb{E}_S[W^*_{k,S}]} \geq 2$ leads to (19). □

3.2 The largest connected component

This test rejects for large values of the size (number of nodes) of the largest connected component in $\mathcal{G}$, which we denoted $\mathcal{C}_{\max}$.

3.2.1 Subcritical regime

We first study that test in the subcritical regime where $\limsup \lambda_0 < 1$. Define

$$I_\lambda = \lambda - 1 - \log(\lambda) . \tag{20}$$

Theorem 2 (Subcritical largest connected component test)**.** *Assume that* $\limsup \lambda_0 < 1$*,* $\log\log(N) = o(\log n)$ *and* $I_{\lambda_0}^{-1} \log(N) \to \infty$*. The largest connected component test is asymptotically powerful when either* $\liminf \lambda_1 > 1$ *or*

$$\lambda_0 \leq \lambda_1 e^{1-\lambda_1} \textit{ for } n \textit{ large enough} \quad \textit{and} \quad \liminf \frac{I_{\lambda_0}}{\lambda_0 + I_{\lambda_1} - \lambda_0 e^{I_{\lambda_1}}} \frac{\log(n)}{\log(N)} > 1 . \tag{21}$$

If we further assume that $n^2 = o(N)$*, then the largest connected component test is asymptotically powerless when* $\lambda_1 < 1$ *for all* n *and*

$$\lambda_0 \geq \lambda_1 e^{1-\lambda_1} \textit{ for } n \textit{ large enough} \quad \textit{or} \quad \limsup \frac{I_{\lambda_0}}{\lambda_0 + I_{\lambda_1} - \lambda_0 e^{I_{\lambda_1}}} \frac{\log(n)}{\log(N)} < 1 . \tag{22}$$

If we assume that both λ_0 and λ_1 go to zero, then Condition (21) is equivalent to

$$\liminf \frac{I_{\lambda_0}}{I_{\lambda_1}} \frac{\log(n)}{\log(N)} > 1 , \tag{23}$$

which corresponds to the optimal detection boundary in this setting, as shown in Theorem 4.

The technical hypothesis $\log\log(N) = o(\log n)$ is only used for convenience when analyzing the critical behavior $\lambda_1 \to 1$. The condition $I_{\lambda_0}^{-1} \log(N) \to \infty$ implies that λ_0 can only converge to zero slower than any power of N. Although it is possible to analyze the test in the very sparse setting where λ_0 goes to zero polynomially fast, we did not do so to keep the exposition focused on the more interesting regimes.

Proof of Theorem 2. That the test is powerful when $\liminf \lambda_1 > 1$ derives from the well-known phase transition phenomenon of Erdös-Rényi graphs.

Lemma 8. *Let* $\mathcal{C}_m$ *denote a largest connected component of* $\mathbb{G}(m, \lambda/m)$ *and assume that* $\lambda \in (0, \infty)$ *is fixed. Then, in probability,*

$$|\mathcal{C}_m| \sim \begin{cases} I_\lambda^{-1} \log m, & \textit{if } \lambda < 1 ; \\ (1-\eta_\lambda) m, & \textit{if } \lambda > 1 . \end{cases}$$

Proof. When $\lambda > 1$, see (Van der Hofstad, 2012, Th. 4.8) — the result is actually contained in Lemma 5. When $\lambda < 1$, see (Van der Hofstad, 2012, Th. 4.4, Th. 4.5). $\square$

Hence, under the null with $\limsup \lambda_0 < 1$, the largest connected component of $\mathcal{G}$ is of order $\log(N)$ with probability going to one. Under the alternative H_S with $\liminf \lambda_1 > 1$, the graph $\mathcal{G}_S$ contains a giant connected component whose size of order n with probability going to one. Recalling that $\log(N) = o(n)$ allows us to conclude.

Now suppose that (21) holds. We assume that the sequence λ_1 is always smaller or equal to 1, that $I_{\lambda_1}^{-1} = O(\log(n)/\log(N))$ and that $\log(I_{\lambda_1}^{-1} \vee 1) = o(\log n)$, meaning that λ_1 does not converge too fast to 1. We may do so while keeping Condition (21) true because the distribution of $|\mathcal{C}_{\max}|$ under $\mathbb{P}_S$ is stochastically increasing with λ_1, because $\limsup \lambda_0 < 1$, $I_{\lambda_1} + \lambda_0 - \lambda_0 e^{I_{\lambda_1}} \sim I_{\lambda_1}(1-\lambda_0)$ for $\lambda_1 \to 1$, and because $\log\log(N) = o\log(n)$.

By hypothesis (21), there exists a constant $c' > 0$, such that

$$\tau := \liminf \frac{I_{\lambda_0}\log(n)}{(I_{\lambda_1} + \lambda_0 - \lambda_0 e^{I_{\lambda_1}})\log(N)} \geq 1 + c' \ .$$

To upper-bound the size of $\mathcal{C}_{\max}$ under $\mathbb{P}_0$, we use the following.

Lemma 9. *Let $\mathcal{C}_m$ denote a largest connected component of $\mathbb{G}(m, \lambda/m)$ and assume that $\lambda < 1$ for all m and $\log[I_\lambda^{-1} \vee 1] = o(\log(m))$. Then, for any sequence u_m satisfying*

$$\liminf \frac{u_m I_\lambda}{\log m} > 1 \ ,$$

we have

$$\mathbb{P}(|\mathcal{C}_m| \geq u_m) = o(1) \ .$$

Proof. This lemma is a slightly modified version of (Van der Hofstad, 2012, Th. 4.4), the main difference difference being that λ was fixed in the original statement. Details are omitted. $\square$

Define $c = (c' \wedge 1)/4$. Applying Lemma 9, $|\mathcal{C}_{\max}| \leq t_0 := I_{\lambda_0}^{-1}\log(N)(1+c)$, with probability going to one under $\mathbb{P}_0$.

We now need to lower-bound the size of $\mathcal{C}_{\max}$ under $\mathbb{P}_S$. Define

$$\begin{aligned} k_0 &= (1-c)\log(n)\left[I_{\lambda_1} + \lambda_0 - \lambda_0 e^{I_{\lambda_1}}\right]^{-1} \ , \quad k = \lceil k_0 \rceil \ , \\ q_0 &= (1-c)\log(n)\frac{1 - \lambda_0 e^{I_{\lambda_1}}}{I_{\lambda_1} + \lambda_0 - \lambda_0 e^{I_{\lambda_1}}} \ , \quad q = \lfloor q_0 \rfloor \ . \end{aligned}$$

The denominator of k_0 is positive since $\lambda_0 e^{I_{\lambda_1}} \leq 1$ and

$$I_{\lambda_1} + \lambda_0 - \lambda_0 e^{I_{\lambda_1}} \geq I_{\lambda_1} + e^{-I_{\lambda_1}}\left(1 - e^{I_{\lambda_1}}\right) = I_{e^{-I_{\lambda_1}}} > 0 \ . \tag{24}$$

We note that $k = O(\log n)$, unless the denominator of k_0 goes to zero, which is only possible when I_{λ_1} goes to zero (implying $\lambda_1 \to 1$), in which case

$$k \sim \log(n)[I_{\lambda_1}(1-\lambda_0)]^{-1} = O\left[I_{\lambda_1}^{-1} \vee 1\right]\log(n) = O\left[\log(N)\right] \ , \tag{25}$$

since, in this case, (21) implies that $I_{\lambda_1}^{-1} = O(\log(n)/\log(N))$, and $\limsup \lambda_0 < 1$ by assumption. So (25) holds in any case.

We shall prove that among the connected components of $\mathcal{G}_S$ of size larger than q, there exists at least one component whose size in $\mathcal{G}$ is larger than k. By definition of c, we have $\liminf k/t_0 \geq \tau(1-c)/(1+c) \geq (1+c')(1-c)/(1+c) > 1$, and the connected component test is therefore powerful. The main arguments rely on the second moment method and on the comparison between cluster sizes and branching processes. Before that, recall that $t_0 \to \infty$, so that $\log(n) I^{-1}_{e^{-I_{\lambda_1}}} \succ k_0 \to \infty$, which in turn implies $I_{\lambda_1} = o(\log(n))$.

Lemma 10. *Fix any $c > 0$. Consider the distribution $\mathbb{G}(m, \lambda/m)$ and assume that λ satisfies*

$$\limsup \lambda \leq 1, \qquad \log\left[I_\lambda^{-1} \vee 1\right] = o\left(\log(m)\right) , \qquad I_\lambda^{-1} \log m \to \infty .$$

For any sequence $q = a\log(m)$ with $a \leq I_\lambda^{-1}(1-c)$, let $Z_{\geq q}$ denote the number of nodes belonging to a connected component whose size is larger than q. With probability going to one, we have

$$Z_{\geq q} \geq m^{1-aI_\lambda - o(1)} . \tag{26}$$

Proof. This lemma is a simple extension of the second method argument (Equations (4.3.34) and (4.3.35)) in the proof of (Van der Hofstad, 2012, Th. 4.5), where λ is fixed, while here it may vary with m, and in particular, may converge to 1. We leave the details to the reader. □

Observe that

$$\frac{q}{(1-c)I_{\lambda_1}^{-1}\log(n)} \leq \frac{I_{\lambda_1} - \lambda_0 I_{\lambda_1} e^{I_{\lambda_1}}}{I_{\lambda_1} + \lambda_0 - \lambda_0 e^{I_{\lambda_1}}} \leq 1 - \lambda_0 \frac{1 - e^{I_{\lambda_1}} + I_{\lambda_1} e^{I_{\lambda_1}}}{I_{\lambda_1} + \lambda_0 - \lambda_0 e^{I_{\lambda_1}}} \leq 1 ,$$

using the fact that $xe^x - e^x + 1 \geq 0$ for any $x \geq 0$. Thus, we can apply Lemma 10 to $\mathcal{G}_S$. And by Lemma 9, the largest connected component of $\mathcal{G}_S$ has size smaller than $2I_{\lambda_1}^{-1}\log(n)$ with probability tending to one. Hence, $\mathcal{G}_S$ contains more than

$$\frac{n^{1+o(1)}e^{-qI_{\lambda_1}}}{2I_{\lambda_1}^{-1}\log n} = ne^{-qI_{\lambda_1} - o(\log(n))}$$

connected components of size larger than q. (We used the fact that $\log(I_{\lambda_1}^{-1} \vee 1) = o(\log n)$.) If $k_0 - q_0 \leq 1$, then applying Lemma 10 to $q+2$ (instead of q) allows us to conclude that there exists a connected component of size at least k. This is why we assume in the following that $\liminf k_0 - q_0 > 1$. By definition of k_0 and q_0, $k_0 - q_0 \geq 1$, implies that

$$\log(n)\lambda_0 \geq \frac{1}{1-c} e^{-I_{\lambda_1}} \left(I_{\lambda_1} + \lambda_0 - \lambda_0 e^{I_{\lambda_1}}\right) \geq \frac{1}{1-c} e^{-I_{\lambda_1}} I_{e^{-I_{\lambda_1}}}$$

by (24). Thus, $\liminf k_0 - q_0 > 1$ implies that for n large enough $\log(n)\lambda_0 \geq \lambda_1 I_{\lambda_1} e$ and consequently

$$I_{\lambda_0} \leq O(1) - \log(\lambda_0) \leq o\left(\log(n)\right) + I_{\lambda_1} + \log\left[I^{-1}_{e^{-I_{\lambda_1}}}\right] = o(\log(n)) \tag{27}$$

since $I_{\lambda_1} = o(\log(n))$, $-\log(I_{\lambda_1}) \leq o(\log(n))$ and $I^{-1}_{e^{-I_\lambda}} = O\left[(e^{-I_\lambda} - 1)^{-2}\right] = O\left[I_\lambda^{-2}\right]$.

Let $\{\mathcal{C}_S^{(i)},\ i \in \mathcal{I}\}$ denote the collection of connected component of size larger than q in $\mathcal{G}_S$. For any such component $\mathcal{C}_S^{(i)}$, we extract any subconnected component $\tilde{\mathcal{C}}_S^{(i)}$ of size q. Recall that, with probability going to one:

$$|\mathcal{I}| \geq n^{1-o(1)}e^{-qI_{\lambda_1}} . \tag{28}$$

For any node x, let denote $\mathcal{C}(x)$ the connected component of x in $\mathcal{G}$, and let $\mathcal{C}_{-S}(x)$ denote the connected component of x in the graph $\mathcal{G}_{-S}$ where all the edges in $\mathcal{G}_S$ have been removed. Then, let

$$U_i := \bigcup_{x \in \tilde{\mathcal{C}}_S^{(i)}} \mathcal{C}_{-S}(x) \ , \ i \in \mathcal{I} \ ; \qquad V = \sum_{i \in \mathcal{I}} \mathbb{1}_{\{|U_i| \geq k\}} \ .$$

Since $V \geq 1$ implies that the largest connected component of $\mathcal{G}$ is larger than k, it suffices to prove that V is larger than one with probability going to one. Observe that conditionally to $|\mathcal{I}|$, the distribution of $(|U_i|,\ i \in \mathcal{I})$ is independent of $\mathcal{G}_S$. Again, we use a second moment method based on a stochastic comparison between connected components and binomial branching processes.

Lemma 11 (Upper bound on the cluster sizes)**.** *Consider the distribution $\mathbb{G}(m, \lambda/m)$ and a collection $\mathcal{J}$ of nodes. For each $k \geq |\mathcal{J}|$,*

$$\mathbb{P}\left[|\cup_{x \in \mathcal{J}} \mathcal{C}(x)| \geq k\right] \leq \mathbb{P}_{m,\lambda/m}\left[T_1 + \ldots + T_{|\mathcal{J}|} \geq k\right] \ ,$$

where $T_1, T_2, \ldots$ denote the total progenies of i.i.d. binomial branching processes with parameters m and λ/m. For each $|\mathcal{J}| \leq k \leq m$,

$$\mathbb{P}\left[|\cup_{x \in \mathcal{J}} C(x)| \geq k\right] \geq \mathbb{P}_{m-k,\lambda/m}\left[T_1 + \ldots + T_{|\mathcal{J}|} \geq k\right] \ ,$$

where $T_1, T_2, \ldots$ denote the total progenies of i.i.d. binomial branching processes with parameters $m - k$ and λ/m.

Lemma 11 is a slightly modified version of (Van der Hofstad, 2012, Th. 4.2 and 4.3), the only difference difference being that $|\mathcal{J}| = 1$ in the original statement. The proof is left to the reader. The following result is proved in (Van der Hofstad, 2012, Sec. 3.5).

Lemma 12 (Law of the total progeny)**.** *Let $T_1, \ldots, T_r$ denote the total progenies of r i.i.d. branching processes with offspring distribution X. Then,*

$$\mathbb{P}\left[T_1 + \ldots + T_r = k\right] = \frac{r}{k}\, \mathbb{P}\left[X_1 + \ldots + X_k = k - r\right] \ ,$$

where (X_i), $i = 1, \ldots, k$ are i.i.d. copies of of X.

Relying on these three lemmas, we control the conditional expectation and variance of V.

Lemma 13. *The following bounds hold*

$$\begin{aligned} \mathbb{P}_S[|U_i| \geq k] \quad &\geq \quad \left(\frac{k}{k-q}\right)^{k-q} e^{-\lambda_0 q - I_{\lambda_0}(k-q)} n^{-o(1)} \ , \\ \mathrm{Var}_S[V|\mathcal{G}_S] \quad &\leq \quad |\mathcal{I}|\, \mathbb{P}_S[|U_i| \geq k] + \frac{|\mathcal{I}|^2 q^2}{N}\, \mathbb{E}_S[|U_i| \mathbb{1}_{\{U_i \geq k\}}] \ , &(29) \\ \mathbb{P}_S[|U_i| \geq k] \quad &\leq \quad \mathbb{E}_S[|U_i| \mathbb{1}_{\{U_i \geq k\}}] \leq \left(\frac{k}{k-q}\right)^{k-q} e^{-\lambda_0 q - I_{\lambda_0}(k-q)} n^{o(1)} \ . &(30) \end{aligned}$$

Before proceeding to the proof of Lemma 13, we finish proving that $V \geq 1$ with probability going to one. Let define $\mu_k := \left(\frac{k}{k-q}\right)^{k-q} e^{-\lambda_0 q - I_{\lambda_0}(k-q)}$. Applying Chebyshev inequality, we derive from Lemma 13

$$V \geq |\mathcal{I}| \mu_k n^{-o(1)} - O_{\mathbb{P}_S}\left[\left(|\mathcal{I}|\mu_k\right)^{1/2} n^{o(1)}\right] - O_{\mathbb{P}_S}\left[|\mathcal{I}|(\mu_k/N)^{1/2} n^{o(1)}\right] \ .$$

In order to conclude, we only to need to prove that $|\mathcal{I}|\mu_k \geq n^{c-o(1)}$ since $(|\mathcal{I}|\mu_k)^{1/2} / |\mathcal{I}|(\mu_k/N)^{1/2} = \sqrt{N/|\mathcal{I}|} \geq 1$. First, we consider $|\mathcal{I}|\mu_k$. Relying on (28), we derive

$$
\begin{aligned}
|\mathcal{I}|\mu_k &\geq n^{1-o(1)} \left(\frac{k}{k-q}\right)^{k-q} e^{-\lambda_0 q - qI_{\lambda_1} - I_{\lambda_0}(k-q)} \\
&\geq n^{1-o(1)} \left(\frac{k_0}{k_0-q_0}\right)^{k_0-q_0} e^{-\lambda_0 q_0 - q_0 I_{\lambda_1} - I_{\lambda_0}(k_0-q_0) - 2I_{\lambda_0}} \\
&\geq n^{1-o(1)} \lambda_0^{-(k_0-q_0)} e^{-\lambda_0 q_0 - k_0 I_{\lambda_1} - I_{\lambda_0}(k_0-q_0)} \\
&\geq n^{1-o(1)} e^{-k_0\lambda_0 - k_0 I_{\lambda_1}} e^{k_0-q_0} \\
&\geq n^{1-o(1)} \exp\left[-k_0\left(\lambda_0 + I_{\lambda_1} - \lambda_0 e^{I_{\lambda_1}}\right)\right] = n^{c-o(1)} ,
\end{aligned}
$$

where we use (27) and $\frac{k_0}{k_0-q_0} = \lambda_0^{-1} e^{-I_{\lambda_1}}$ in the third line, the definition $I_{\lambda_0} = \lambda_0 - \log(\lambda_0) - 1$ in the fourth line, and the definitions of k_0 and q_0 in the last line.

Proof of Lemma 13. Consider any subset $\mathcal{J}$ of node of size q. The distribution $|U_i| = |\bigcup_{x \in \tilde{\mathcal{C}}_S^{(i)}} \mathcal{C}_{-S}(x)|$ is stochastically dominated by the distribution of $Z := |\bigcup_{x \in \mathcal{J}} \mathcal{C}(x)|$ under the null hypothesis. Let T_q be sum of the total progenies of q independent binomial branching processes with parameters $N-n+q-k$ and p_0. By Lemma 11, we derive

$$
\mathbb{P}_S[|U_i| \geq k] \geq \mathbb{P}_0[Z \geq k] \geq \mathbb{P}_{N-n+q-k,p_0}[T_q \geq k] \geq \mathbb{P}_{N-n+q-k,p_0}[T_q = k] .
$$

Let $X_1, X_2, \ldots$ denote independent binomial random variables with parameters $N-n+q-k$ and p_0. Relying on Lemma 12 and the lower bound $\binom{s}{r} \geq \frac{(s-r)^r}{r!} \geq (re)^{-1}\left(\frac{(s-r)e}{r}\right)^r$, we derive

$$
\begin{aligned}
\mathbb{P}_{N-n+q-k,p_0}[T_q = k] &= \frac{q}{k}\, \mathbb{P}_{N-n+q-k,p_0}[X_1 + \ldots + X_k = k-q] \\
&= \frac{q}{k} \binom{k(N-n+q-k)}{k-q} p_0^{k-q} (1-p_0)^{k(N-n+q-k)-k+q} \\
&\succ \frac{q}{k^2} \left[\frac{ek(N-n-2(q-k))}{k-q}\right]^{k-q} \left(\frac{\lambda_0}{N}\right)^{k-q} e^{-\lambda_0 k - kO(n/N)} \\
&\succ \frac{q}{k^2} e^{-I_{\lambda_0}(k-q)} e^{-\lambda_0 q} \left(\frac{k}{k-q}\right)^{k-q} e^{-kO(n/N)} \\
&\succ \left(\frac{k}{k-q}\right)^{k-q} e^{-\lambda_0 q - I_{\lambda_0}(k-q)} n^{o(1)} ,
\end{aligned}
$$

where (25) with $n\log(N)/N = o(\log(n))$ in the last line.

Let us now prove (30). The first inequality is Markov's. For the second, by Lemma 11, U_i is stochastically dominated by $\tilde{T}_q$, the sum of the total progenies of q independent binomial branching processes with parameters N and p_0, so that

$$
\mathbb{E}_S\left[|U_i| \mathbb{1}_{\{U_i \geq k\}}\right] = \sum_{r=k}^{N} \mathbb{P}_S[U_i \geq r] \leq \sum_{r=k}^{\infty} \mathbb{P}_{N,p_0}[\tilde{T}_q \geq r] = \sum_{r=k}^{\infty} r\, \mathbb{P}_{N,p_0}[\tilde{T}_q = r] .
$$

We use Lemma 12 to control the deviation of $\tilde{T}_q$. Below $X_1, X_2, \ldots$ denote independent binomial

random variables with parameter N and p_0.

$$\begin{aligned}\sum_{r=k}^{\infty} r\,\mathbb{P}_{N,p_0}[\tilde{T}_q = r] &\leq \sum_{r=k}^{\infty} r\frac{q}{r}\,\mathbb{P}_{N,p_0}[X_1+\ldots+X_r = r-q] \\ &\leq \sum_{r=k}^{\infty} q\exp\left[-NrH_{p_0}\left(\frac{r-q}{Nr}\right)\right]\ , \end{aligned} \tag{31}$$

by Chernoff inequality since

$$\frac{r-q}{Nr} \geq \frac{k-q}{Nk} \geq \frac{k_0-q_0}{Nk_0} = \frac{\lambda_0 e^{I_{\lambda_1}}}{N} > \frac{\lambda_0}{N} = p_0\ .$$

By Lemma 2, $H_{p_0}(a) \geq a\log(a/p_0) - a + p_0$. Thus, we arrive at

$$\begin{aligned}\mathbb{E}_S\left[|U_i|\mathbb{1}_{\{U_i\geq k\}}\right] &\leq \sum_{r=k}^{\infty} q\exp\left[-(r-q)\log\left(\frac{r-q}{r\lambda_0}\right) + r - q - r\lambda_0\right] \\ &\leq q\sum_{r=k}^{\infty}\exp[A_r]\ ;\quad A_r := -(r-q)I_{\lambda_0} - q\lambda_0 - (r-q)\log\left(\frac{r-q}{r}\right)\ . \end{aligned} \tag{32}$$

Differentiating the function A_r with respect to r, we get

$$\begin{aligned}\frac{dA_r}{dr} &= -I_{\lambda_0} - \log\left(\frac{r-q}{r}\right) - 1 + \frac{r-q}{r} \leq -I_{\lambda_0} - \log\left(\frac{k-q}{k}\right) - 1 + \frac{k-q}{k} \\ &\leq -I_{\lambda_0} - \log\left(\frac{k_0-q_0}{k_0}\right) - 1 + \frac{k_0-q_0}{k_0} = -\lambda_0 - I_{\lambda_1} + \lambda_0 e^{I_{\lambda_1}}\ ,\end{aligned}$$

which is negative as argued below the definition of k. Consequently, A_r is a decreasing function of r. Define r_1 as the smallest integer such that $\log((r-q)/r) \geq -I_{\lambda_0}/2$. Since $\limsup \lambda_0 < 1$, it follows $r_1 = O(q)$. Coming back to (32), we derive

$$\begin{aligned}\mathbb{E}_S\left[|U_i|\mathbb{1}_{\{U_i\geq k\}}\right] &\leq q(r_1-k)_+\exp[A_k] + q\sum_{r=r_1}^{\infty}\exp[A_r] \\ &\leq qe^{A_k}\left[(r_1-k)_+ + \sum_{r=r_1}^{\infty} e^{-(r-k)\left[I_{\lambda_0}-\log((r-q)/r)\right]}\right] \\ &\leq qe^{A_k}\left[(r_1-k)_+ + \sum_{r=r_1}^{\infty} e^{-(r-k)I_{\lambda_0}/2}\right] \\ &\leq e^{A_k}O(k^2)\ ,\end{aligned} \tag{33}$$

since $\limsup \lambda_0 < 1$. From (25), we know that $k = O(\log(N)) = n^{o(1)}$, which allows us to prove (30).

Turning to the proof of (29), we have the decomposition

$$\begin{aligned}\mathrm{Var}_S[V|\mathcal{G}_S] &\leq |\mathcal{I}|\,\mathbb{P}_S[U_i \geq k] + \sum_{i\neq i'\in\mathcal{I}}\left\{\mathbb{P}_S\left[|U_i|\geq k,\, |U_{i'}|\geq k\right] - \mathbb{P}_S^2\left[|U_i|\geq k\right]\right\} \\ &\leq |\mathcal{I}|\,\mathbb{P}_S[U_i\geq k] + |\mathcal{I}|^2\,\mathbb{P}_S\left[|U_i|\geq k\ ,\ U_i\cap U_{i'}\neq\emptyset\right] \\ &\quad +|\mathcal{I}|^2\left\{\mathbb{P}_S\left[|U_i|\geq k,\, |U_{i'}|\geq k\ , U_i\cap U_{i'}=\emptyset\right] - \mathbb{P}_S^2\left[|U_i|\geq k\right]\right\}\ .\end{aligned} \tag{34}$$

The last term is nonpositive. Indeed,

$$\begin{aligned}
&\mathbb{P}_S\left[|U_i| \geq k,\ |U_{i'}| \geq k\ , U_i \cap U_{i'} = \emptyset\right] - \mathbb{P}_S^2\left[|U_i| \geq k\right] \\
&= \sum_{r=k}^{N} \mathbb{P}_S\left[|U_i| = r\right] \left(\mathbb{P}_S\left[|U_{i'}| \geq k,\ U_i \cap U_{i'} = \emptyset \,\middle|\, |U_i| = r\right] - \mathbb{P}_S\left[|U_{i'}| \geq k\right]\right) \\
&\leq \sum_{r=k}^{N} \mathbb{P}_S\left[|U_i| = r\right] \left(\mathbb{P}_S\left[|U_{i'}| \geq k \,\middle|\, U_i \cap U_{i'} = \emptyset\ ,\ |U_i| = r\right] - \mathbb{P}_S\left[|U_{i'}| \geq k\right]\right)\ ,
\end{aligned}$$

where the last difference is negative, as the graph is now smaller once we condition on $|U_i| \geq 1$ and $U_i \cap U_{i'} = \emptyset$. Consider the second term in (34):

$$\mathbb{P}_S\left[|U_i| \geq k\ ,\ U_i \cap U_{i'} \neq \emptyset\right] = \sum_{r=k}^{N} \mathbb{P}_S[|U_i| = r]\, \mathbb{P}_S[U_i \cap U_{i'} \neq \emptyset \,|\, |U_i| = r]\ .$$

By symmetry and a union bound, we derive

$$\mathbb{P}_S[U_i \cap U_{i'} \neq \emptyset \,|\, |U_i| = r] \leq q^2\, \mathbb{P}_S[y \in \mathcal{C}_{-S}(x) \,|\, |U_i| = r]\ ,$$

for some $x \in \tilde{\mathcal{C}}_S^{(i)}$ and $y \in \tilde{\mathcal{C}}_S^{(i')}$. Since the graph $\mathcal{G}_{-S}$ is not symmetric, the probability that a fixed node z belongs to $\mathcal{C}_{-S}(x)$ conditionally to $|\mathcal{C}_{-S}(x)|$ is smaller for $z \in S \setminus \{i\}$ than for $z \in S^c$. It follows that

$$\mathbb{P}_S[y \in \mathcal{C}_{-S}(x) \,|\, |U_i| = r] \leq \mathbb{E}_S\left[\frac{|\mathcal{C}_{-S}(x)| - 1}{N-1} \,\middle|\, |U_i| = r\right]\ .$$

Since $|\mathcal{C}_{-S}(x)| \leq r$, we conclude

$$\mathbb{P}_S\left[|U_i| \geq k\ ,\ U_i \cap U_{i'} \neq \emptyset\right] \leq \sum_{r=k}^{N} \mathbb{P}_S[|U_i| = r] \frac{q^2 r}{N} = \frac{q^2}{N}\, \mathbb{E}_S[|U_i| \mathbb{1}_{\{U_i \geq k\}}]\ .$$

□

Let us continue with the proof of Theorem 2, now assuming that $\lambda_1 < 1$, that Condition (22) holds, and that $n^2 = o(N)$. We assume in the sequel that $I_{\lambda_1} \leq -\log(\lambda_0)$, meaning that λ_1 is not too small. We may do so while keeping Condition (22) true, because the distribution of $|\mathcal{C}_{\max}|$ under $\mathbb{P}_S$ is increasing with respect to λ_1 and because for $I_{\lambda_1} = -\log(\lambda_0)$, (22) is equivalent to $\limsup \log(n)/\log(N) < 1$, which is always true since $n^2 = o(N)$. Similarly, we assume that $I_{\lambda_1} = o(\log(n))$ while keeping Condition (22) true since for I_{λ_1} going to infinity, (22) is equivalent to $\limsup \frac{I_{\lambda_0} \log(n)}{I_{\lambda_1} \log(N)} < 1$ and since $I_{\lambda_0}^{-1} \log(N) \to \infty$. By Condition (22), there exists a constant $c > 0$ such that

$$\limsup\ \frac{I_{\lambda_0}}{\lambda_0 + I_{\lambda_1} - \lambda_0 e^{I_{\lambda_1}}} \frac{\log(n)}{\log(N)} < 1 - c\ . \tag{35}$$

We shall prove that with probability $\mathbb{P}_S$ going to one, the largest connected component of $\mathcal{G}$ does not intersect S. As the distribution of the statistic under the alternative dominates the distribution under the null, this will imply that the largest connected component test is asymptotically powerless. Denote $\mathcal{A}$ the event

$$\mathcal{A} := \{\text{For all } (x, y) \in S, \text{there is no path between } x \text{ and } y \text{ with all other nodes in } S^c\}\ .$$

For any subset T, denote $\mathcal{C}_T(x)$ the connected component of x in $\mathcal{G}_T$, and recall that $\mathcal{C}(x)$ is a shorthand for $\mathcal{C}_{\mathcal{V}}(x)$. By symmetry, we have

$$\mathbb{P}_S[\mathcal{A}^c] \leq n^2\, \mathbb{P}_0[y \in \mathcal{C}_{-S}(x)] \leq \mathbb{P}_0[y \in \mathcal{C}(x)] \ ,$$

since the probability of the edges outside $\mathcal{G}_S$ under $\mathbb{P}_S$ is the same as under $\mathbb{P}_0$. Again, by symmetry

$$\mathbb{P}_0[y \in \mathcal{C}(x)] = \mathbb{E}_0[\mathbb{P}_0[y \in C(x)] \,|\, |\mathcal{C}(x)|] \leq \mathbb{E}_0\left[\frac{|\mathcal{C}(x)|}{N-1}\right] \leq \frac{1}{(N-1)(1-\lambda_0)} \ ,$$

as the expected size of a cluster is dominated by the expected progeny of a branching process with parameters N and p_0 (Lemma 11) and the expected progeny of a subcritical branching process having mean offspring $\mu < 1$ is $(1-\mu)^{-1}$ (Van der Hofstad, 2012, Th. 3.5). Thus,

$$\mathbb{P}_S[\mathcal{A}^c] = O(n^2/N) = o(1) \ . \tag{36}$$

Define

$$k := (1-c)^{1/2} \log(N) I_{\lambda_0}^{-1} \ . \tag{37}$$

Since $\limsup \lambda_0 < 1$ and since $\log\log(N) = o[\log(n)]$, it follows that $k \asymp \log(N) = n^{o(1)}$. By Lemma 10, $|\mathcal{C}_{\max}|$ is larger or equal to k with probability $\mathbb{P}_S$ (and $\mathbb{P}_0$) going to one. Thus, it suffices to prove that $\mathbb{P}_S[\vee_{x\in S}|\mathcal{C}(x)| \geq k] \to 0$. Observe that

$$\mathbb{P}_S[\vee_{x\in S}|\mathcal{C}(x)| \geq k] \leq n\, \mathbb{P}_S\left[\{|C(x)| \geq k\} \cap \mathcal{A}\right] + \mathbb{P}_S[\mathcal{A}^c] \ ,$$

so that, by (36), we only need to prove that $n\, \mathbb{P}_S\left[\{|C(x)| \geq k\} \cap \mathcal{A}\right] = o(1)$. Under the event $\mathcal{A}$, $\mathcal{C}(x) \cap S$ is exactly the connected component $\mathcal{C}_S(x)$ of x in $\mathcal{G}_S$. Furthermore, $\mathcal{C}(x)$ is the union of $\mathcal{C}_{-S}(y)$ over $y \in \mathcal{C}_S(x)$. Consequently, we have the decomposition

$$\mathbb{P}_S\left[\{|\mathcal{C}(x)| \geq k\} \cap \mathcal{A}\right] \quad \leq \quad \mathbb{P}_S[|\mathcal{C}_S(x)| \geq k] + \sum_{q=1}^{k-1} \mathbb{P}_S[|\mathcal{C}_S(x)| = q]\, \mathbb{P}_S\left[\mathcal{B}_q \,|\, |\mathcal{C}_S(x)| = q\right] \ ,$$

where $\mathcal{B}_q := \{|\cup_{y\in\mathcal{C}_S(x)} \mathcal{C}_{-S}(y)| \geq k\}$. By Lemma 11, the distribution of $|\mathcal{C}_S(x)|$ is stochastically dominated by the total progeny distribution of a binomial branching process with parameters $(n, \lambda_1/n)$. Denote by $\mathcal{J}$ any set of nodes of size q. Since, conditionally to $|\mathcal{C}_S(x)| = q$, the event $\mathcal{B}_q$ is increasing and only depends on the edges outside $\mathcal{G}_S$, we have

$$\mathbb{P}_S\left[\mathcal{B}_q | \, |\mathcal{C}_S(x)| = q\right] \leq \mathbb{P}_0\left[|\cup_{y\in\mathcal{J}} \mathcal{C}(y)| \geq k\right] \ ,$$

which is in turn, by Lemma 11, smaller than the probability that the total progeny of q independent branching processes with parameters $(N, \lambda_0/N)$ is larger than k. Relying on the law of the total progeny of branching processes (Lemma 12) and Lemma 11, we get

$$\begin{aligned}
\mathbb{P}_S[|\mathcal{C}_S(x)| = q] \quad &\leq \quad \frac{1}{q}\, \mathbb{P}\left[\mathrm{Bin}(nq, \lambda_1/n) = q-1\right] \ , \\
\mathbb{P}_S\left[\mathcal{B}_q \,|\, |\mathcal{C}_S(x)| = q\right] \quad &\leq \quad \sum_{r=k}^{\infty} \frac{q}{r}\, \mathbb{P}\left[\mathrm{Bin}(Nr, \lambda_0/N) = r-q\right] \ .
\end{aligned}$$

Working out the density of the binomial random variable, we derive

$$\mathbb{P}_S[|\mathcal{C}_S(x)| = q] \quad \leq \quad \binom{nq}{q-1} p_1^{q-1}(1-p_1)^{nq-q+1} \prec \frac{1}{\lambda_1} e^{-I_{\lambda_1} q} \ ,$$

and for $q \le (1-\lambda_0)k$, we get

$$\mathbb{P}_S\left[\mathcal{B}_q \,|\, |\mathcal{C}_S(x)| = q\right] \quad \le \quad \frac{q}{k} \exp\left[-NkH_{p_0}\left(\frac{k-q}{Nk}\right)\right] \ ,$$

which is exacly the term (31), which has been proved in (33) to be smaller than

$$O(k^2)\left(\frac{k-q}{k}\right)^{k-q} e^{-(k-q)I_{\lambda_0} - q\lambda_0} \ .$$

Let define

$$B_\ell := e^{-I_{\lambda_1}\ell - \ell\lambda_0 - (k-\ell)I_{\lambda_0}}\left(\frac{k}{k-\ell}\right)^{k-\ell}$$

Gathering all these bounds, we get

$$\begin{aligned}
\mathbb{P}_S\left[\{|C(x)| \ge k\} \cap \mathcal{A}\right] \quad &\prec \quad \frac{e^{-I_{\lambda_1}k}}{\lambda_1} + \sum_{q=\lceil(1-\lambda_0)k\rceil}^{k-1} \frac{e^{-I_{\lambda_1}q}}{\lambda_1} + O\left(\frac{k^2}{\lambda_1}\right)\sum_{q=1}^{\lfloor(1-\lambda_0)k\rfloor} B_q \\
&\prec \quad \frac{k^3}{\lambda_1}\left[e^{-I_{\lambda_1}(1-\lambda_0)k} + \bigvee_{q=1}^{k} B_q\right] \prec n^{o(1)} \sup_{q\in[0;k]} B_q \ ,
\end{aligned}$$

where we observe that $e^{-I_{\lambda_1}(1-\lambda_0)k} = B_{(1-\lambda_0)k}$ and we use $k = n^{o(1)}$ and $I_{\lambda_1} = o(\log(n))$. By differentiating $\log(B_q)$ as a function of q, we obtain the maximum

$$\sup_{q\in[0;k]} B_q \le \begin{cases} e^{-kI_{\lambda_0}} \ , & \text{if} \quad \lambda_0 e^{I_{\lambda_1}} > 1 \ ; \\ e^{-I_{\lambda_1}k}\exp\left[\lambda_0 k(e^{I_{\lambda_1}} - 1)\right] \ , & \text{else} \ . \end{cases}$$

Recall that we assume $\lambda_0 e^{I_{\lambda_1}} \le 1$ so that

$$\mathbb{P}_S\left[\{|C(x)| \ge k\} \cap \mathcal{A}\right] \quad \prec \quad n^{o(1)}\frac{1}{\lambda_1}\exp\left[-k\left\{\lambda_0 + I_{\lambda_1} - \lambda_0 e^{I_{\lambda_1}}\right\}\right] \prec n^{-(1-c)^{-1/2}+o(1)} \ ,$$

by definition (37) of k and Condition (35). We conclude that $n\,\mathbb{P}_S\left[\{|C(x)| \ge k\} \cap \mathcal{A}\right] = o(1)$. □

3.2.2 Supercritical regime

We now briefly discuss the behavior of the largest connected component test in the supercritical regime where $\liminf \lambda_0 > 1$. When $\lambda_0 - \log N \to \infty$, the graph $\mathcal{G}$ is connected with probability tending to one under the null and under any alternative (Van der Hofstad, 2012, Th. 5.5), which renders the test completely useless. We focus on the case where λ_0 is fixed for the sake of simplicity. In that regime, we find that, in that case, the test performs roughly as well as the total degree test — compare Proposition 1.

Proposition 3 (Supercritical largest connected component test)**.** *The largest connected component test is asymptotically powerful when $\lambda_1 > \lambda_0 > 1$ are fixed and $n^2/N \to \infty$.*

Proof. We keep the same notation. Under $\mathbb{P}_0$, we have $|\mathcal{C}_{\max}| = (1-\eta_{\lambda_0})N + O(\sqrt{N})$ (Van der Hofstad, 2012, Th. 4.16). Hereafter, assume that we are under $\mathbb{P}_S$. Then, by the same token, $|\mathcal{C}^{S^c}_{\max}| = (1-\eta_{\lambda_0})(N-n) + O(\sqrt{N-n})$ and $|\mathcal{C}^{S}_{\max}| = (1-\eta_{\lambda_1})n + O(\sqrt{n})$. Given $\mathcal{G}_S$ and $\mathcal{G}_{S^c}$,

the probability that $\mathcal{C}_{\max}^{S^c}$ and $\mathcal{C}_{\max}^{S}$ are connected in $\mathcal{G}$ is equal to $1-(1-p_0)^{|\mathcal{C}_{\max}^{S^c}|\,|\mathcal{C}_{\max}^{S}|} \to 1$ in probability, since $p_0|\mathcal{C}_{\max}^{S^c}|\,|\mathcal{C}_{\max}^{S}| \asymp n$ in probability. Hence, with probability tending to one,

$$\begin{aligned}|\mathcal{C}_{\max}| &\geq |\mathcal{C}_{\max}^{S^c}| + |\mathcal{C}_{\max}^{S}| \\ &= (1-\eta_{\lambda_0})(N-n) + O(\sqrt{N}) + (1-\eta_{\lambda_1})n + O(\sqrt{n}) \\ &= (1-\eta_{\lambda_0})N + (\eta_{\lambda_0}-\eta_{\lambda_1})n + O(\sqrt{N})\ ,\end{aligned}$$

with $\eta_{\lambda_0} - \eta_{\lambda_1} > 0$ since $\lambda_1 > \lambda_0 > 1$ and η_λ is strictly decreasing. Hence, because $n \gg \sqrt{N}$ by assumption, the test that rejects when $|\mathcal{C}_{\max}| \geq (1-\eta_{\lambda_0})N + \frac{1}{2}(\eta_{\lambda_0}-\eta_{\lambda_1})n$. □

When $\lambda_0 > 1$ is fixed, the largest connected component is of size $|\mathcal{C}_{\max}|$ satisfying

$$\frac{|\mathcal{C}_{\max}| - (1-\eta_{\lambda_0})N}{\sqrt{N}} \to \mathcal{N}(0,1), \quad \text{under } \mathbb{P}_0\ ,$$

by (Van der Hofstad, 2012, Th. 4.16), while $|\mathcal{C}_{\max}|$ increases by at most n under the alternative, so the test is powerless when $n = o(\sqrt{N})$.

3.3 The number of k-trees

We consider the test that rejects for large values of N_k^{tree}, the number of subtrees of size k. This test will partially bridge the gap in constants between what the broad scan test and largest connected component test can achieve in the regime where λ_0 is constant. Recall the definition of I_λ in (20).

Theorem 3. *Assume that λ_1 and λ_0 are both fixed, with $0 < \sqrt{\lambda_0/e} < \lambda_1 < 1$, and that*

$$\limsup \frac{\log(N/n^2)}{\log n} < \frac{I_{\frac{\lambda_0}{\lambda_1 e}} - I_{\sqrt{\frac{\lambda_0}{e}}}}{(1-\frac{\lambda_0}{\lambda_1 e})I_{\frac{\sqrt{\lambda_1}}{e}}}\ . \tag{38}$$

Then there is a constant $c > 0$ such that the test based on N_k^{tree} with $k = c\log n$ is asymptotically powerful.

Thus, even in the supercritical Poissonian regime with $1 < \lambda_0 < e$, there exist subcritical communities $\lambda_1 < 1$ that are asymptotically detectable with probability going to one. The condition $\lambda_1 > \sqrt{\lambda_0/e}$ will be shown to be minimal in Theorem 5. Condition (38) essentially requires that n^2/N does not converge too fast to zero. In particular, when $n = N^\kappa$, (38) translates into an upper bound on κ. We show later in Theorem 5 that such an upper bound is unavoidable, for when κ is too small, no test is asymptotically powerful. Nevertheless, Condition (38) is in all likelihood not optimal.

Proof of Theorem 3. We first compute the expectation of N_k^{tree} under $\mathbb{P}_0$ using Cayley's formula. Since $k^2 = o(n) = o(N)$ and $k \to \infty$, we derive

$$\begin{aligned}\mathbb{E}_0[N_k^{\text{tree}}] &= \sum_{|C|=k} \mathbb{P}_0[\mathcal{G}_C \text{ is a tree}] \\ &= \binom{N}{k} k^{k-2} p_0^{k-1} (1-p_0)^{k^{(2)}-k+1} \\ &\sim N(\lambda_0 e)^k \frac{1}{\sqrt{2\pi}k^{5/2}\lambda_0}\ ,\end{aligned}$$

where we used the fact any k-tree has exactly $k-1$ edges. The last line comes from an application of Stirling's formula. We then bound the variance of N_k^{tree} under $\mathbb{P}_0$ in the following lemma, whose lengthy proof is postponed to Section 7.3.

Lemma 14. *When $\lambda_0 < e$, we have*

$$\mathrm{Var}_0[N_k^{\text{tree}}] \prec \frac{N}{k\lambda_0}(e\lambda_0)^k e^{2k\sqrt{\lambda_0/e}} \ .$$

By Chebyshev's inequality, under $\mathbb{P}_0$,

$$N_k^{\text{tree}} = \mathbb{E}_0\, N_k^{\text{tree}} + O\big(\,\mathrm{Var}_0(N_k^{\text{tree}})\big)^{1/2} \ .$$

Fix $S \subset \mathcal{V}$ of size $|S| = n$, and let q be an integer between 1 and k chosen later. We let N_{k,S^c}^{tree} denote the number of k-trees in $\mathcal{G}_{S^c}$, and let $N_{k,S,q}^{\text{tree}}$ as the number of subsets C of size k such that $|C \cap S| = q$ and both $\mathcal{G}_{C\cap S}$ and $\mathcal{G}_C$ are trees. We have $N_k^{\text{tree}} \geq N_{k,S^c}^{\text{tree}} + N_{k,S,q}^{\text{tree}}$. Therefore, by Chebyshev's inequality, under $\mathbb{P}_S$

$$N_k^{\text{tree}} \geq \mathbb{E}_S\left(N_{k,S^c}^{\text{tree}}\right) + \mathbb{E}_S\left(N_{k,S,q}^{\text{tree}}\right) + O\big(\,\mathrm{Var}_S(N_{k,S^c}^{\text{tree}})\big)^{1/2} + O\big(\,\mathrm{Var}_S(N_{k,S,q}^{\text{tree}})\big)^{1/2} \ .$$

Noting that $\mathcal{G}_{S^c} \sim \mathbb{G}(N-n, p_0)$, and letting $\lambda_0' = (N-n)p_0$, Lemma 14 implies that

$$\mathrm{Var}_S[N_{k,S^c}^{\text{tree}}] \prec \frac{N-n}{k\lambda_0'}(e\lambda_0')^k e^{2k\sqrt{\lambda_0'/e}} \sim \frac{N}{k\lambda_0}(e\lambda_0)^k e^{2k\sqrt{\lambda_0/e}} \ ,$$

because $nk = o(N)$. Thus, we only need to show that, for a careful choice of q,

$$\mathbb{E}_S[N_{k,S,q}^{\text{tree}}] \gg \mathbb{E}_0[\tilde{N}_k^T] - \mathbb{E}_S[N_{k,S^c}^{\text{tree}}] \ , \tag{39}$$

$$\mathbb{E}_S^2[N_{k,S,q}^{\text{tree}}] \gg \mathrm{Var}_S[N_{k,S,q}^{\text{tree}}] \ , \tag{40}$$

$$\mathbb{E}_S^2[N_{k,S,q}^{\text{tree}}] \gg \frac{N}{k\lambda_0}(e\lambda_0)^k e^{2k\sqrt{\lambda_0/e}} \succ \mathrm{Var}_0[N_k^{\text{tree}}] \ . \tag{41}$$

From now on, let $q = k - \lfloor \frac{\lambda_0}{\lambda_1 e} k \rfloor$.

We use the following lemma, whose lengthy proof is postponed to Section 7.4.

Lemma 15. *When $q = k - \lfloor \frac{\lambda_0}{\lambda_1 e} k \rfloor$, we have*

$$\mathbb{E}_S[N_{k,S,q}^{\text{tree}}] \succ n\frac{\lambda_1^{k-1}e^{2k-q}}{k^3} \succ n(e\lambda_1)^k e^{\frac{\lambda_0}{\lambda_1 e}k}\frac{1}{\lambda_1 k^3} \ , \tag{42}$$

and

$$\mathrm{Var}_S[N_{k,S,q}^{\text{tree}}] \prec nk^2\lambda_1^{2k-q-1}e^{4k-2q}e^{2\frac{\sqrt{\lambda_1}}{e}q} + \frac{k^7n^2}{N}\lambda_1^{2k-2}\lambda_0 e^{4k-2q} \ . \tag{43}$$

We first prove (39), bounding

$$\begin{aligned}
\mathbb{E}_0[N_k^{\text{tree}}] - \mathbb{E}_S[\tilde{N}_{k,S^c}^T] &= \left(\binom{N}{k} - \binom{N-n}{k}\right) k^{k-2}p_0^{k-1}(1-p_0)^{k^{(2)}-k+1} \\
&\leq \left(N^k - (N-n-k)^k\right)\frac{k^{k-2}}{k!}\left(\frac{\lambda_0}{N}\right)^{k-1} \\
&\prec n(\lambda_0 e)^k k^{-5/2} \ ,
\end{aligned}$$

since $[1-(n+k)/N]^k = 1 + kn/N + o(kn/N)$ by the fact that $k = o(n)$ and $kn = o(N)$. We also used Stirling's formula again. Using this bound together with (42), we derive

$$\frac{\mathbb{E}_S[N^{\mathrm{tree}}_{k,S,q}]}{\mathbb{E}_0[N^{\mathrm{tree}}_k] - \mathbb{E}_S[\tilde{N}^T_{k,S^c}]} \succ \frac{\lambda_0}{k^{1/2}\lambda_1}\left(\frac{\lambda_1}{\lambda_0}\right)^k e^{k\frac{\lambda_0}{\lambda_1 e}} = \frac{\lambda_0}{k^{1/2}\lambda_1}\exp\left[kI_{\frac{\lambda_0}{\lambda_1 e}}\right] \to \infty \ ,$$

since λ_0 and λ_1 are fixed such that $\lambda_0/\lambda_1 e < 1$, implying that $I_{\frac{\lambda_0}{\lambda_1 e}} > 0$ is fixed.

Second, we prove (40). Using (42) and (43), we have

$$\begin{aligned}\frac{\mathrm{Var}_S[N^{\mathrm{tree}}_{k,S,q}]}{\mathbb{E}^2_S[N^{\mathrm{tree}}_{k,S,q}]} &\prec \frac{k^8}{n}\lambda_1^{-q}e^{2\frac{\sqrt{\lambda_1}}{e}q} + \frac{k^{13}}{N} \\ &\prec \frac{k^8}{n}\exp\left[2k\big(1-\frac{\lambda_0}{\lambda_1 e}\big)I_{\frac{\sqrt{\lambda_1}}{e}}\right] + \frac{k^{13}}{N} \ ,\end{aligned}$$

and the RHS goes to 0 as long as

$$\limsup \frac{k}{\log(n)} < \frac{1}{2\big(1-\frac{\lambda_0}{\lambda_1 e}\big)I_{\frac{\sqrt{\lambda_1}}{e}}} \ .$$

Finally, we prove (41). Using Lemma 14 and (42), we have

$$\begin{aligned}\frac{\mathrm{Var}_0[N^{\mathrm{tree}}_k]}{\mathbb{E}^2_S[N^{\mathrm{tree}}_{k,S,q}]} &\prec \frac{Nk^5}{n^2}\left(\frac{\lambda_1^2}{e\lambda_0}\right)^k e^{2k\sqrt{\frac{\lambda_0}{e}}-4k+2q} \\ &\prec \frac{Nk^5}{n^2}\exp\left[2k\big(I_{\sqrt{\frac{\lambda_0}{e}}} - I_{\frac{\lambda_0}{\lambda_1 e}}\big)\right] \ .\end{aligned}$$

Note that $I_{\sqrt{\frac{\lambda_0}{e}}} - I_{\frac{\lambda_0}{\lambda_1 e}} < 0$ is fixed, since our assumptions imply that $\frac{\lambda_0}{\lambda_1 e} < \sqrt{\frac{\lambda_0}{e}} < 1$ and the function I_λ is decreasing on $(0,1)$. Thus, the RHS above goes to 0 as long as

$$\liminf \frac{k}{\log(N/n^2)} > \frac{1}{2\big(I_{\frac{\lambda_0}{\lambda_1 e}} - I_{\sqrt{\frac{\lambda_0}{e}}}\big)} \ .$$

$\square$

3.4 The number of triangles

We recall that this test is based on the number T of triangles in $\mathcal{G}$. This is an emblematic test among those based on counting patterns, as it is the simplest and the least costly to compute. As such, the number of triangles in a graph is an important topological characteristic, with applications in the study of real-life networks. For example, Maslov et al. (2004) use the number of triangles to quantify the amount of clustering in the Internet.

Proposition 4. *The triangle test is asymptotically powerful if either*

$$\limsup \lambda_0 < \infty \quad \textit{and} \quad \lambda_1 \to \infty \ ; \tag{44}$$

or

$$\liminf \lambda_0 > 0 \quad , \quad \lambda_0 < N/n \quad \textit{and} \quad \frac{\lambda_1^2}{\lambda_0} \gg 1 \vee \left(\frac{\lambda_0}{\sqrt{N}}\right)^{2/3} \ . \tag{45}$$

When λ_0 and λ_1 are fixed, T converges in distribution towards a Poisson distribution with parameter $\lambda_0^3/6$ under the null and $(\lambda_0^3+\lambda_1^3)/6$ under the alternative hypothesis. In particular, the test is not *asymptotically powerless if*

$$\limsup \lambda_0 < \infty \qquad \textit{and} \qquad \liminf \lambda_1 > 0 \ . \tag{46}$$

Proof of Proposition 4. Let T be the number of triangles in $\mathcal{G}$. For $S \subset \mathcal{V}$, let T_S denote the number of triangles in $\mathcal{G}_S$. We have $T \geq T_{S^\mathsf{c}} + T_S$.

The following result is based on (Bollobás, 2001, Th. 4.1, 4.10). We use it multiple times below without explicitly saying so.

Lemma 16. *Let T_m be the number of triangles in $\mathbb{G}(m, \lambda/m)$. If λ is fixed, then $T_m \Rightarrow \text{Poisson}(\lambda^3/6)$. If $\lambda \to \infty$ with $\log\lambda = o(\log m)$, then $\frac{T_m-\mu}{\sqrt{\mu}} \Rightarrow \mathcal{N}(0,1)$ where $\mu := \mathbb{E}\, T = \binom{m}{3}\left(\frac{\lambda}{m}\right)^3 \sim \frac{\lambda^3}{6}$.*

Assume that (44) holds. Applying Lemma 16, $T \to 0$ under $\mathbb{P}_0$, while $T \geq T_S \to \infty$ under $\mathbb{P}_S$. (For the latter, we use the fact that T is stochastically increasing in λ_1.)

Assume that λ_0 and λ_1 are fixed. Applying Lemma 16, $T \Rightarrow \text{Poisson}(\lambda_0^3/6)$ under $\mathbb{P}_0$, while $T \Rightarrow \text{Poisson}(\lambda_0^3/6)$ under $\mathbb{P}_0$, while under $\mathbb{P}_S$, $T_{S^\mathsf{c}} + T_S \Rightarrow \text{Poisson}((\lambda_0^3+\lambda_1)^3/6)$ since $T_{S^\mathsf{c}} \sim \mathbb{G}(N-n, p_0)$ and $T_S \sim \mathbb{G}(n, p_1)$ are independent, and $n = o(N)$. Define $T_{S,S^\mathsf{c}} := T - T_S - T_{S^\mathsf{c}}$ as the number of triangles in $\mathcal{G}$ with nodes both in S and S^c. We have

$$E_S\left[T_{S,S^\mathsf{c}}\right] \leq N^2 n p_0^3 + n^2 N p_1 p_0^2 \leq \frac{n}{N}\lambda_0^3 + \frac{n}{N}\lambda_1\lambda_0^2 = o(1) \ ,$$

so that $T_{S,S^\mathsf{c}} = o_{\mathbb{P}_S}(1)$, and by Slutsky's theorem, $T \Rightarrow \text{Poisson}((\lambda_0^3+\lambda_1)^3/6)$ under $\mathbb{P}_S$.

Assume that (46) holds. By considering a subsequence if needed, we may assume that $\lambda_0 < \infty$ is fixed. And since T is stochastically increasing in λ_1 under the alternative, we may assume that $\lambda_1 > 0$ is fixed. We have proved above that $T \Rightarrow \text{Poisson}(\lambda_0^3/6)$ under $\mathbb{P}_0$, $T \Rightarrow \text{Poisson}(\lambda_0^3/6+\lambda_1^3/6)$ the alternative; hence the test $\{T \geq 1\}$ has risk

$$\mathbb{P}_0(T \geq 1) + \mathbb{P}_S(T_S = 0) \to 1 - e^{-\lambda_0^3/6} + e^{-\lambda_0^3/6-\lambda_1^3/6} < 1 \ .$$

Finally, assume that (45) holds. Using Chebyshev's inequality, to prove that the test based on T is powerful it suffices to show that

$$\frac{\mathbb{E}_S\, T - \mathbb{E}_0\, T}{\sqrt{\text{Var}_S(T) \vee \text{Var}_0(T)}} \to \infty \ . \tag{47}$$

Straightforward calculations show that $\mathbb{E}_0\, T = \binom{N}{3}p_0^3$, and

$$\begin{aligned} \text{Var}_0(T) &= \left[3(N-3)(1-p_0)p_0^2 + (1-p_0^3)\right]\binom{N}{3}p_0^3 \\ &\asymp N^4 p_0^5 + N^3 p_0^3 \ . \end{aligned}$$

And carefully counting the number of triplets with 2 or 3 vertices in S gives

$$\mathbb{E}_S\, T = \binom{N-n}{3}p_0^3 + \binom{n}{1}\binom{N-n}{2}p_0^3 + \binom{n}{2}\binom{N-n}{1}p_0^2 p_1 + \binom{n}{3}p_1^3 \ ,$$

while counting pairs of triplets with a certain number of vertices in S, shared or not, we arrive at the rough estimate

$$\text{Var}_S(T) \asymp N^4 p_0^5 + n^2 N^2 p_0^4 p_1 + n^3 N p_0^2 p_1^3 + n^4 p_1^5 + \mathbb{E}_S\, T \ .$$

Note that

$$\begin{aligned}\mathbb{E}_S T - \mathbb{E}_0 T &= \binom{n}{2}\binom{N-n}{1} p_0^2(p_1 - p_0) + \binom{n}{3}(p_1^3 - p_0^3) \\ &\asymp n^2(p_1 - p_0)\left[Np_0^2 + np_1^2\right] \\ &\succ Nn^2p_0^2p_1 + n^3p_1^3\ , \end{aligned} \tag{48}$$

since by condition (45), $np_0 \leq 1$ and $np_1 = \lambda_1 \gg 1$. and

$$\mathrm{Var}_0(T) \prec \mathrm{Var}_S(T) \asymp N^4p_0^5 + n^2N^2p_0^4p_1 + n^3Np_0^2p_1^3 + n^4p_1^5 + N^3p_0^3 + n^3p_1^3\ .$$

We only need to prove that the square root of this last expression is much smaller than (48). Since $(np_1)^2 \gg Np_0$ and $np_1 \to \infty$, we first derive that

$$n^3Np_0^2p_1^3 + n^4p_1^5 + N^3p_0^3 + n^3p_1^3 = o\left[(np_1)^6\right]\ .$$

Similarly, we get $n^2N^2p_0^4p_1 = o\left[n^4p_1^2N^2p_0^4\right]$. Finally, (45) entails that $\lambda_1^2 \gg \lambda_0(\lambda_0/\sqrt{N})^{2/3}$ which is equivalent to $N^4p_0^5 = o\left[(np_1)^6\right]$. □

4 Information theoretic lower bounds

In this section we state and prove lower bounds on the risk of any test whatsoever. In most cases, we find sufficient conditions under which the null and alternative hypotheses merge asymptotically, meaning that all tests are asymptotically powerless. In other cases, we find sufficient conditions under which no test is asymptotically powerful.

To derive lower bounds, it is standard to reduce a composite hypothesis to a simple hypothesis. This is done by putting a prior on the set of distributions that define the hypothesis. In our setting, we assume that p_0 is known so that the null hypothesis is simple, corresponding to the Erdös-Rényi model $\mathbb{G}(N, p_0)$. The alternative $H_1 := \bigcup_{|S|=n} H_S$ is composite and 'parametrized' by subsets of nodes of size n. We choose as prior the uniform distribution over these subsets, leading to the simple hypothesis $\bar{H}_1$ comprising of $\mathbb{G}(N, p_0; n, p_1)$ defined earlier. The corresponding risk for H_0 versus $\bar{H}_1$ is

$$\bar{\gamma}_N(\phi) = \mathbb{P}_0(\phi = 1) + \frac{1}{\binom{N}{n}} \sum_{|S|=n} \mathbb{P}_S(\phi = 0)\ .$$

Note that $\gamma_N(\phi) \geq \bar{\gamma}_N(\phi)$ for any test ϕ. Our choice of prior was guided by invariance considerations: the problem is invariant with respect to a relabeling of the nodes. In our setting, this implies that $\gamma_N^* = \bar{\gamma}_N^*$, or equivalently, that there exists a test invariant with respect to permutation of the nodes that minimizes the worst-case risk (Lehmann and Romano, 2005, Lem. 8.4.1). Once we have a simple versus simple hypothesis testing problem, we can express the risk in closed form using the corresponding likelihood ratio. Let $\bar{\mathbb{P}}_1$ denote the distribution of $\mathbf{W}$ under $\bar{H}_1$, meaning $\mathbb{G}(N, p_0; n, p_1)$. The likelihood ratio for testing $\mathbb{P}_0$ versus $\bar{\mathbb{P}}_1$ is

$$L = \frac{1}{\binom{N}{n}} \sum_{|S|=n} L_S\ , \tag{49}$$

where L_S be the likelihood for testing $\mathbb{P}_0$ versus $\mathbb{P}_S$. Then the test $\phi^* = \{L > 1\}$ is the unique test that minimizes $\bar{\gamma}_N$, and

$$\bar{\gamma}_N(\phi^*) = \bar{\gamma}_N^* = 1 - \frac{1}{2}\,\mathbb{E}_0\,|L - 1|\ .$$

For each subset $S \subset \mathcal{V}$ of size n, let Γ_S be an event, i.e., a subset of adjacency matrices, and define the truncated likelihood as

$$\tilde{L} = \frac{1}{\binom{N}{n}} \sum_{|S|=n} L_S \, \mathbb{1}_{\Gamma_S} \ . \tag{50}$$

We have

$$\begin{aligned}
\mathbb{E}_0 \, |L - 1| &\leq \mathbb{E}_0 \, |\tilde{L} - 1| + \mathbb{E}_0 (L - \tilde{L}) \\
&\leq \sqrt{\mathbb{E}_0[\tilde{L}^2] - 1 + 2(1 - \mathbb{E}_0[\tilde{L}])} + (1 - \mathbb{E}_0[\tilde{L}]) \ ,
\end{aligned}$$

using the Cauchy-Schwarz inequality and the fact that $\mathbb{E}_0 \, L = 1$ since it is a likelihood. Hence, for all tests to be asymptotically powerless, it suffices that $\mathbb{E}_0[\tilde{L}^2] \leq 1 + o(1)$ and $\mathbb{E}_0[\tilde{L}] \geq 1 + o(1)$. Note that

$$\mathbb{E}_0[\tilde{L}] = \frac{1}{\binom{N}{n}} \sum_{|S|=n} \mathbb{P}_S(\Gamma_S) \ .$$

In all our examples, $\mathbb{P}_S(\Gamma_S)$ is independent of S, so that $\mathbb{E}_0[\tilde{L}] \geq 1 + o(1)$ is equivalent to $\mathbb{P}_S(\Gamma_S) \to 1$.

4.1 All tests are asymptotically powerless

We start with some sufficient conditions under which all tests are asymptotically powerless. Recall α in (7) and ζ in (12). We require that $\zeta \to 0$ below to prevent the total degree test from having any power (see Proposition 1).

Theorem 4. *Assume that $\zeta \to 0$. Then all tests are asymptotically powerless in either of the following situations:*

$$\lambda_0 \to 0, \quad \lambda_1 \to 0, \quad \limsup \frac{I_{\lambda_0}}{I_{\lambda_1}} \frac{\log n}{\log N} < 1 \ ; \tag{51}$$

$$0 < \liminf \lambda_0 \leq \limsup \lambda_0 < \infty, \quad \lambda_1 \to 0 \ ; \tag{52}$$

$$\lambda_0 \to \infty \textit{ with } \alpha \to 0, \quad \limsup \lambda_1 < 1 \ ; \tag{53}$$

$$0 < \liminf \alpha \leq \limsup \alpha < 1, \quad \limsup \ (1 - \alpha) \sup_{k = n/u_N}^{n} \frac{\mathbb{E}_S[W^*_{k,S}]}{k} < 1 \ . \tag{54}$$

We recall here the first few steps that we took in (Arias-Castro and Verzelen, 2012) to derive analogous lower bounds in the denser regime where $\liminf \alpha \geq 1$. We start with some general identities. We have

$$L_S := \exp(\theta W_S - \Lambda(\theta) n^{(2)}) \ , \tag{55}$$

with

$$\theta := \theta_{p_1}, \quad \theta_q := \log \left(\frac{q(1 - p_0)}{p_0(1 - q)} \right) \ , \tag{56}$$

and

$$\Lambda(\theta) := \log(1 - p_0 + p_0 e^{\theta}) \ ,$$

which is the moment generating function of Bern(p_0).

In all cases, the events Γ_S satisfy

$$\Gamma_S \subset \bigcap_{k > k_{\min}} \{ W_T \leq w_k, \ \forall T \subset S \text{ such that } |T| = k \} \ , \tag{57}$$

where $k_{\min}$ and w_k vary according to the specific setting.

To prove that $\mathbb{E}_0 \tilde{L}^2 \leq 1 + o(1)$, we proceed as follows. We have

$$\begin{aligned}
\mathbb{E}_0 \left[\tilde{L}^2\right] &= \frac{1}{\binom{N}{n}^2} \sum_{|S_1|=n} \sum_{|S_2|=n} \mathbb{E}_0 \left(L_{S_1} L_{S_2} \mathbb{1}_{\Gamma_{S_1}} \mathbb{1}_{\Gamma_{S_2}}\right) \\
&= \frac{1}{\binom{N}{n}^2} \sum_{|S_1|=n} \sum_{|S_2|=n} \mathbb{E}_0 \left[\exp\left(\theta(W_{S_1} + W_{S_2}) - 2\Lambda(\theta) n^{(2)}\right) \mathbb{1}_{\Gamma_{S_1} \cap \Gamma_{S_2}}\right] .
\end{aligned}$$

Define

$$W_{S\times T} = \frac{1}{2} \sum_{i \in S, j \in T} W_{i,j} \ ,$$

and note that $W_S = W_{S\times S}$. We use the decomposition

$$W_{S_1} + W_{S_2} = W_{S_1 \times (S_1 \setminus S_2)} + W_{S_2 \times (S_2 \setminus S_1)} + 2W_{S_1 \cap S_2} \ , \tag{58}$$

and the independence of the random variables on the RHS of (58), to get

$$\mathbb{E}_0 \left(\exp\left(\theta(W_{S_1} + W_{S_2}) - 2\Lambda(\theta) n^{(2)}\right) \mathbb{1}_{\Gamma_{S_1} \cap \Gamma_{S_2}}\right) \leq \mathrm{I} \cdot \mathrm{II} \cdot \mathrm{III} \ , \tag{59}$$

where $K = |S_1 \cap S_2|$,

$$\mathrm{I} := \mathbb{E}_0 \left[\exp\left(\theta W_{S_1 \times (S_1 \setminus S_2)} - \frac{\Lambda(\theta)}{2}(n-K)(n+K-1)\right)\right] = 1 \ ,$$

$$\mathrm{II} := \mathbb{E}_0 \left[\exp\left(\theta W_{S_2 \times (S_2 \setminus S_1)} - \frac{\Lambda(\theta)}{2}(n-K)(n+K-1)\right)\right] = 1 \ ,$$

$$\mathrm{III} := \mathbb{E}_0 \left[\exp\left(2\theta W_{S_1 \cap S_2} - 2\Lambda(\theta) K^{(2)}\right) \mathbb{1}_{\Gamma_{S_1} \cap \Gamma_{S_2}}\right] \ .$$

The first two equalities are due to the fact that the likelihood integrates to one.

Assuming that $\zeta \to 0$, we prove that all tests are asymptotically powerless in the following settings:

$$\limsup \lambda_0 < \infty, \quad \lambda_1^2 = o(\lambda_0) \ ; \tag{60}$$

$$\lambda_0 \to 0, \quad \lambda_1 \to 0, \quad \limsup \frac{I_{\lambda_0} \log(n)}{I_{\lambda_1} \log(N)} < 1, \quad n^2 = o(N) \ ; \tag{61}$$

$$\limsup \lambda_1 < 1, \quad \lambda_0 \to \infty \ , \quad \limsup \alpha < 1 \ ; \tag{62}$$

$$\liminf \lambda_1 \geq 1, \quad 0 < \liminf \alpha \leq \limsup \alpha < 1, \quad \limsup \, (1-\alpha) \sup_{k=n/u_N}^{n} \frac{\mathbb{E}_S[W_{k,S}^*]}{k} < 1 \ . \tag{63}$$

This implies Theorem 4. Indeed, (60) includes (52). Assume that (51) holds. Consider any subsequence n^2/N converging to $x \in \mathbb{R}^+ \cup \{\infty\}$. If $x = 0$, then (61) holds. If $x \neq 0$, then $\zeta \to 0$ implies that $(\lambda_1 - \lambda_0 n/N)^2/\lambda_0 = o(1)$. If, in addition, $\lambda_1 \geq 2\lambda_0 n/N$, this implies that $\lambda_1^2/\lambda_0 = o(1)$. If, otherwise, $\lambda_1 \leq 2\lambda_0 n/N$, then $\lambda_1^2/\lambda_0 \leq 4\lambda_0 (n/N)^2 = o(1)$ since $\lambda_0 = o(1)$. Thus, in both cases, (60) holds. Finally, (62) includes (53) and also (54) when $\limsup \lambda_1 < 1$, while (63) includes (54) when $\liminf \lambda_1 \geq 1$. We note that (63) implies that $\limsup \lambda_1 < \infty$ because of (15).

4.1.1 Proof of Theorem 4 under (60)

The arguments here are very similar to those used in (Arias-Castro and Verzelen, 2012), except for the choice of events Γ_S. Define

$$\Gamma_S := \{\mathcal{G}_S \text{ is a forest}\} \ .$$

When Γ_S holds, for any $T \subset S$, $\mathcal{G}_T$ is also a forest, and since any forest $\mathcal{F}$ with k nodes and t connected components (therefore all trees) has exactly $k - t \leq k$ edges, we have $W_T \leq |T|$. Hence, (57) holds with $w_k := k$.

Lemma 17. *$\mathbb{P}_S(\Gamma_S)$ is independent of S of size n, and $\mathbb{P}_S(\Gamma_S) = 1 + o(1)$.*

Proof. The expected number of loops of size k in $\mathcal{G}_S$ under $\mathbb{P}_S$ is equal to

$$\frac{n!}{(n-k)!2k}\, p_1^k \leq \frac{\lambda_1^k}{2k} \ . \tag{64}$$

Summing (64) over k, we see that the expected number of loops in $\mathcal{G}_S$ under $\mathbb{P}_S$ is smaller than $\lambda_1^3/(1-\lambda_1) = o(1)$. Hence, with probability going to one under $\mathbb{P}_S$, $\mathcal{G}_S$ has no loops and is therefore a forest. □

In order to conclude, we only need to prove that $\mathbb{E}_0[\tilde{L}^2] \leq 1 + o(1)$. We start from (59) and we recall that $K = |S_1 \cap S_2|$. We take $k_{\min}$ as the largest integer k satisfying.

$$\frac{2}{k-3} \geq \frac{p_1^2(1-p_0)}{p_0(1-p_1)^2} \ ,$$

with the convention $2/0 = \infty$, so that $k_{\min} \geq 3$. Let $q_k = 2/(k-1)$. Recall that $\rho = n/(N-n)$ and define $k_0 = \lceil bn\rho \rceil$, where $b \to \infty$ satisfies $b^2\zeta \to 0$.

- When $K \leq k_{\min}$, we will use the obvious bound:

$$\mathrm{III} \ \leq \ \mathbb{E}_0 \exp\left(2\theta W_{S_1\cap S_2} - 2\Lambda(\theta)K^{(2)}\right) = \exp\left(\Delta K^{(2)}\right),$$

  where

$$\Delta := \Lambda(2\theta) - 2\Lambda(\theta) = \log\left(1 + \frac{(p_1-p_0)^2}{p_0(1-p_0)}\right) . \tag{65}$$

- When $K > k_{\min}$, we use a different bound. Noting that $\Gamma_{S_1} \cap \Gamma_{S_2} \subset \{W_{S_1\cap S_2} \leq w_K\}$, for any $\xi \in (0, 2\theta)$, we have

$$\begin{aligned} \mathrm{III} \ &\leq \ \mathbb{E}_0\left[\exp\left(\xi W_{S_1\cap S_2} + (2\theta-\xi)w_K - 2\Lambda(\theta)K^{(2)}\right) \mathbb{1}_{\{W_{S_1\cap S_2} \leq w_K\}}\right] \\ &\leq \ \mathbb{E}_0\left[\exp\left(\xi W_{S_1\cap S_2} + (2\theta-\xi)w_K - 2\Lambda(\theta)K^{(2)}\right)\right] \ , \end{aligned}$$

  so that

$$\mathrm{III} \leq \exp\left(\Delta_K K^{(2)}\right),$$

  where

$$\Delta_k := \min_{\xi\in[0,2\theta]} \Lambda(\xi) + (2\theta-\xi)q_k - 2\Lambda(\theta) \ . \tag{66}$$

Using the fact that $\mathbb{E}_0[\tilde{L}^2] \leq \mathbb{E}_0[\mathrm{III}]$, we have

$$\begin{aligned}\mathbb{E}_0[\tilde{L}^2] &\leq \mathbb{E}\left[\mathbb{1}_{\{K\leq k_0\}}\exp\left(\Delta K^{(2)}\right)\right] \\ &\qquad + \mathbb{E}\left[\mathbb{1}_{\{k_0+1\leq K\leq k_{\min}\}}\exp\left(\Delta K^{(2)}\right)\right] \\ &\qquad\quad + \mathbb{E}\left[\mathbb{1}_{\{k_{\min}+1\leq K\leq n\}}\exp\left(\Delta_K K^{(2)}\right)\right] \\ &= A_1 + A_2 + A_3 \ ,\end{aligned}$$

where the expectation is with respect to $K \sim \mathrm{Hyp}(N, n, n)$. By Lemma 4, K is stochastically bounded by $\mathrm{Bin}(n, \rho)$. Hence, using Chernoff's bound (see Lemma 1), we have

$$\mathbb{P}(K \geq k) \leq \mathbb{P}(\mathrm{Hyp}(N, n, n) \geq k) \leq \mathbb{P}(\mathrm{Bin}(n, \rho) \geq k) \leq \exp\left(-nH_\rho(k/n)\right) . \tag{67}$$

- When $K \leq k_0$, we proceed as follows. If $k_0 = 1$, we simply have

$$A_1 = \mathbb{P}(K \leq 1) \leq 1 \ .$$

If $k_0 \geq 2$, we use the expression (65) of Δ to derive

$$A_1 \leq \exp\left[\Delta k_0^2\right] \leq \exp\left[O(1)\frac{(p_1-p_0)^2}{p_0(1-p_0)}\frac{b^2n^4}{N^2}\right] = \exp\left[O(b^2\zeta)\right] = 1 + o(1) \ .$$

- When $k_0 + 1 \leq K \leq k_{\min}$, we use (67) and the identity $(1-x)\log(1-x) \geq -x$, to get

$$\begin{aligned} A_2 &\leq \sum_{k=k_0+1}^{k_{\min}} \exp\left[\Delta\frac{k(k-1)}{2} - nH_\rho\left(\frac{k}{n}\right)\right] \\ &\leq \sum_{k=k_0+1}^{k_{\min}} \exp\left[k\left(\Delta\frac{k-1}{2} - \log\left(\frac{k}{n\rho}\right) + 1\right)\right] \ .\end{aligned}$$

The last sum is equal to zero if $k_{\min} \leq k_0$; therefore, assume that $k_{\min} > k_0$. For $a > 0$ fixed, the function $f(x) = ax - \log x$ is decreasing on $(0, 1/a)$ and increasing on $(1/a, \infty)$. Therefore, for $k_0 + 1 \leq k \leq k_{\min}$,

$$\Delta\frac{k-1}{2} - \log\left(\frac{k}{n\rho}\right) \leq \max_{\ell\in\{k_0,k_{\min}\}}\left\{\Delta\frac{\ell-1}{2} - \log\left(\frac{\ell N}{n^2}\right)\right\} \ .$$

We know that $\Delta(k_0 - 1) = o(1)$, so that

$$\Delta\frac{k_0-1}{2} - \log\left(\frac{k_0}{n\rho}\right) \leq o(1) - \log b \to -\infty \ .$$

Therefore, it suffices to show that

$$\frac{k_{\min}-1}{2}\Delta - \log\left(\frac{k_{\min}}{n\rho}\right) \to -\infty \ .$$

If $k_{\min} > 3$, observe that

$$\frac{k_{\min}-1}{2}\Delta \leq \left(1 + \frac{k_{\min}-3}{2}\right)\log\left(1 + \frac{2}{k_{\min}-3}(1+o(1))\right) \leq \frac{3}{2}\log 3 + o(1) \ ,$$

while $\log(k_{\min}/(n\rho)) \geq \log(k_0/(n\rho)) \to \infty$. If we have $k_{\min} = 3$, then we have

$$\Delta - \log\left(\frac{3}{n\rho}\right) \leq \log(p_1^2/p_0) - \log(N/n^2) + O(1) \leq \log\left(\frac{\lambda_1^2}{\lambda_0}\right) + O(1) \to -\infty \ ,$$

because of (60).

- When $k_{\min} < K \leq n$, we need to bound Δ_K. Remember the definition of the entropy function H_q in (9), and that $H(q)$ is short for $H_{p_0}(q)$. It is well-known that H is the Fenchel-Legendre transform of Λ; more specifically, for $q \in (p_0, 1)$,

$$H(q) = \sup_{\theta \geq 0}[q\theta - \Lambda(\theta)] = q\theta_q - \Lambda(\theta_q) \ . \tag{68}$$

Hence, the minimum of $\Lambda(\xi) + (2\theta - \xi)q_k - 2\Lambda(\theta)$ over $\xi > 0$ is achieved at $\xi = \theta_{q_k}$ as soon as $2\theta \geq \theta_{q_k}$. Moreover, by definition of θ in (56), our choice of q_k, and the fact that $k \geq k_{\min}$, we have

$$2\theta - \theta_{q_k} = \log\left(\frac{p_1^2(1-p_0)}{p_0(1-p_1)^2}\frac{2}{k-3}\right) \geq 0 \ .$$

Hence, we have

$$\begin{aligned} \Delta_k &= -H(q_k) + 2\theta q_k - 2\Lambda(\theta) \\ &= -2H_{p_1}(q_k) + H(q_k) \ . \end{aligned} \tag{69}$$

Using the definition of the entropy and the fact that $p_0 = o(1)$, we therefore have

$$\begin{aligned} \Delta_k &= q_k \log\left(\frac{p_1^2}{q_k p_0}\right) + (1-q_k)\log\left(\frac{(1-p_1)^2}{(1-q_k)(1-p_0)}\right) \\ &\leq \frac{2}{k-1}\log\left(\frac{\lambda_1^2 N(k-1)}{2\lambda_0 n^2}\right) + O(1) \ , \end{aligned}$$

where the $O(1)$ is uniform in k. Hence,

$$\begin{aligned} A_3 &\leq \sum_{k=k_{\min}+1}^{n} \exp\left[k\left\{\log\left(\frac{\lambda_1^2}{\lambda_0}\right) + \log\left(\frac{N(k-1)}{2n^2}\right) - \log\left(\frac{Nk}{n^2}\right) + O(1)\right\}\right] \\ &\leq \sum_{k=k_{\min}+1}^{n} \exp\left[k\left\{\log\left(\frac{\lambda_1^2}{\lambda_0}\right) + O(1)\right\}\right] = o(1) \ , \end{aligned}$$

since $\lambda_1^2/\lambda_0 = o(1)$.

This concludes the proof of Theorem 4 under (60).

4.1.2 Proof of Theorem 4 under (61)

Let c be a positive constant that will be chosen small later on. Define

$$f_n := (1+c)\, I_{\lambda_1}^{-1} \log(n) \ .$$

We consider the event

$$\Gamma_S = \{\mathcal{G}_S \text{ is a forest}\} \cap \{|\mathcal{C}_{\max,S}| \leq f_n\} \ .$$

When Γ_S holds, for any $T \subset S$, $\mathcal{G}_T$ is also a forest, with $|T| - W_T$ connected components. Since the size of each connected component is at most f_n, there are at least $\lceil |T|/f_n \rceil$ connected components. Hence, (57) holds with $w_k = k - \lceil \frac{k}{f_n} \rceil$.

Lemma 18. *$\mathbb{P}_S(\Gamma_S)$ is independent of S of size n, and $\mathbb{P}_S(\Gamma_S) = 1 + o(1)$.*

Proof. This is a straightforward consequence of Lemmas 9 and 17. □

To conclude, it suffices to show that $\mathbb{E}_0[\tilde{L}^2] \leq 1 + o(1)$. For this, we will need the following.

Lemma 19. *Let $F_{k,j}$ stand for the number of forests with j trees on k labelled vertices. For any $k \geq 2$ and any $j \leq k$, $F_{k,j} \leq k^{k-2}$.*

Proof. Fix $k \geq 2$. By Cayley's formula, we have $F_{k,1} = k^{k-2}$. Therefore, it suffices to prove that $F_{k,j} \geq F_{k,j+1}$ for all $j \geq 1$. If we take a forest with j trees and erase any of its $k - j$ edges, we obtain a forest with $j + 1$ trees. And there are exactly $\sum_{s \neq t} k_s k_t$ such ways of obtaining a given forest with $j + 1$ trees of sizes $k_1 \leq \cdots \leq k_{j+1}$. Since

$$\sum_{s \neq t} k_s k_t \geq k_1(k - k_1) \geq k - 1 \ ,$$

it follows that $F_{k,j}(k - j) \geq F_{k,j+1}(k - 1)$. Thus, $F_{k,j} \geq F_{k,j+1}$. □

Starting from (59), and using the fact that, under $\Gamma_{S_1} \cap \Gamma_{S_2}$, $\mathcal{G}_{S_1 \cap S_2}$ is a forest with $W_{S_1 \cap S_2} \leq w_K$ edges, we have

$$\mathbb{E}_0\left[\tilde{L}^2\right] \ \leq \ \mathbb{E}_0\left(\exp\left(2\theta W_{S_1 \cap S_2} - 2\Lambda(\theta) K^{(2)}\right) \mathbb{1}_{\{\mathcal{G}_{S_1 \cap S_2} \text{ is a forest}, W_{S_1 \cap S_2} \leq w_K\}}\right) \ .$$

Note that the exponential term is smaller than 1 when $|S_1 \cap S_2| \leq 1$. Recall that $\rho = \frac{m}{N-m}$ and that $\Lambda(\theta) = \log\left[(1 - p_0)/(1 - p_1)\right]$. We derive

$$\begin{aligned}
\mathbb{E}_0\left[\tilde{L}^2\right] - 1 \ &\leq \ \sum_{k=2}^{n} \sum_{i=1}^{w_k} \mathbb{P}\left[K = k, W_{S_1 \cap S_2} = i, \mathcal{G}_{S_1 \cap S_2} \text{ is a forest}\right] \exp\left[2i\theta - 2\Lambda(\theta) k^{(2)}\right] \\
&\leq \ \sum_{k=2}^{n} \sum_{i=1}^{w_k} \binom{n}{k} \rho^k F_{k,k-i} \frac{p_1^{2i}}{p_0^i} \left(\frac{1 - p_0}{1 - p_1}\right)^{2(i - k^{(2)})} \\
&\prec \ \sum_{k=2}^{n} \sum_{i=1}^{w_k} \left(\frac{n^2}{N}\right)^{k-i} \left(\frac{\lambda_1^2}{\lambda_0}\right)^i \frac{F_{k,k-i} \binom{n}{k}}{n^k} \\
&\prec \ \sum_{k=2}^{n} \sum_{i=1}^{w_k} \left(\frac{n^2 e}{N}\right)^{k-i} \left(\frac{\lambda_1^2 e}{\lambda_0}\right)^i \frac{1}{k^2} \\
&\prec \ \sum_{j=1}^{\infty} \left(\frac{n^2 e}{N}\right)^j \sum_{i=1}^{j\lfloor f_n \rfloor} \left(\frac{\lambda_1^2 e}{\lambda_0}\right)^i \frac{1}{(i+j)^2} \\
&\prec \ \sum_{j=1}^{\infty} \left(\frac{n^2 e}{N}\left[1 \vee \frac{\lambda_1^2 e}{\lambda_0}\right]^{f_n}\right)^j \ .
\end{aligned} \tag{70}$$

In the second inequality, we used the fact that K is stochastically bounded by $\text{Bin}(n, \rho)$ (see Lemma 4). In the third inequality, we used the fact that $p_0 < p_1$ and $i \leq w_k < k$, as well as the fact that $n^2 = o(N)$, which implies that $\rho^k \sim (n/N)^k$. In the fourth inequality, we used Lemma 19 and the lower bound $k! \geq (k/e)^k$. The fifth inequality comes from a change of variables and uses

the definition of w_k. When $\lambda_1^2 e \le \lambda_0$, since $n^2 = o(N)$, this sum is $O(n^2/N)$. When $\lambda_1^2 e > \lambda_0$, this sum is equal to

$$\frac{1}{e^{A_n - 1} - 1}, \qquad A_n := \log\left(\frac{N}{n^2}\right) - f_n \log\left(\frac{\lambda_1^2 e}{\lambda_0}\right) . \tag{71}$$

So it suffices to show that $A_n \to \infty$. Since we are working under (61), there is $c > 0$ such that, eventually,

$$\frac{I_{\lambda_0} \log n}{I_{\lambda_1} \log N} \le \frac{1-c}{1+c} .$$

Then, using the fact that $\lambda_0 \vee \lambda_1 = o(1)$, we have

$$\begin{aligned} f_n \log\left(\frac{\lambda_1^2 e}{\lambda_0}\right) &= (1+c)\frac{\log n}{I_{\lambda_1}}\big(2\lambda_1 - 2I_{\lambda_1} + I_{\lambda_0} - \lambda_0\big) \\ &\le -(1+c+o(1))\log(n^2) + (1-c)\log N \\ &\le \log(N/n^2) - c\log(N) , \end{aligned}$$

eventually. This implies that $A_n \ge -1 + c\log N \to \infty$.

This concludes the proof of Theorem 4 under (61).

4.1.3 Proof of Theorem 4 under (62)

Recall that $\rho = n/(N-n)$ and define $k_0 = \lceil bn\rho \rceil$, where $b \to \infty$ satisfies $b^2\zeta \to 0$. Let $k_{\min}$ be the integer part of $1 + \frac{2}{1-\alpha}\big(1 \vee \frac{n^{2-\alpha}}{N^{1-\alpha}}\big)$. Define

$$\Gamma_S = \bigcap_{k=k_{\min}+1}^{n} \{W_T \le w_k,\ \forall T \subset S \text{ such that } |T| = k\} ,$$

where $w_k := k$ here.

Lemma 20. *For any $k > k_{\min}$ and any subset S of size n, we have $\mathbb{P}_S[\Gamma_S] = 1 + o(1)$.*

This takes care of the first moment. In order to conclude, it suffices to control the second moment, specifically, to prove that $\mathbb{E}\big[\tilde{L}^2\big] \le 1 + o(1)$. Arguing as before, we have

$$\begin{aligned} \mathbb{E}_0[\tilde{L}^2] &\le \mathbb{E}\left[\mathbb{1}_{\{K \le k_0\}} \exp\left(\Delta K^{(2)}\right)\right] \\ &\quad + \mathbb{E}\left[\mathbb{1}_{\{k_0+1 \le K \le k_{\min}\}} \exp\left(\Delta K^{(2)}\right)\right] \\ &\quad + \mathbb{E}_0\left[\mathbb{1}_{\{k_0+1 \le K \le k_{\min}\}} \exp\left(2\theta W_{S_1\cap S_2} - 2\Lambda(\theta)K^{(2)}\right) \mathbb{1}_{\{W_{S_1\cap S_2} \le w_K\}}\right] \\ &= A_1 + A_2 + A_3 . \end{aligned}$$

- Arguing exactly as we did before, we have $A_1 = 1 + o(1)$.
- Arguing as before, we also have

$$\begin{aligned} A_2 &\le \sum_{k=k_0+1}^{k_{\min}} \exp\left[k\left(\Delta\frac{k-1}{2} - \log\left(\frac{k}{n\rho}\right) + 1\right)\right] \\ &\le \sum_{k=k_0+1}^{k_{\min}} \exp\left[k\left(1 + o(1) + \max_{\ell \in \{k_0, k_{\min}\}}\left\{\Delta\frac{\ell-1}{2} - \log\left(\frac{\ell N}{n^2}\right)\right\}\right)\right] . \end{aligned}$$

First, we have $\Delta(k_0-1)/2-\log(k_0 N/n^2)\to-\infty$. This is true if $k_0=1$, and when $k_0>1$, we have $N/n^2\le b$, so that

$$\frac{(p_1-p_0)^2}{p_0(1-p_0)}\sim\frac{N^2}{n^4}\zeta=\frac{N^2}{n^4b^2}\,b^2\zeta\to 0\ ,$$

by definition of b, and therefore

$$\Delta\frac{k_0-1}{2}\asymp\frac{N^2}{n^4}\zeta\,\frac{bn^2}{N}\le b^2\zeta\to 0\ .$$

We also have $\Delta(k_{\min}-1)/2-\log(k_{\min}N/n^2)\to-\infty$. To show this, we divide the analysis into two cases. When $N^{1-\alpha}\le n^{2-\alpha}$, this results from

$$\Delta\frac{k_{\min}-1}{2}\le(1+o(1))\frac{n^{2-\alpha}}{(1-\alpha)N^{1-\alpha}}\frac{p_1^2}{p_0}=(1+o(1))\frac{\lambda_1^2}{1-\alpha}=O(1)\ ,$$

together with

$$\log\left(\frac{k_{\min}N}{n^2}\right)\ge\log\left(\frac{2N^\alpha}{(1-\alpha)n^\alpha}\right)\ge\alpha\log(N/n)\to\infty\ , \tag{72}$$

where we used the definition of $k_{\min}$ and the fact that $\lambda_0=(N/n)^\alpha$. When $N^{1-\alpha}\ge n^{2-\alpha}$, this results from

$$\begin{aligned}
\Delta\frac{k_{\min}-1}{2} &\le \frac{1}{2}\lfloor\frac{2}{1-\alpha}\rfloor\log\left(1+\frac{p_1^2}{p_0}\right)+o(1)\\
&\le \frac{1}{2}\lfloor\frac{2}{1-\alpha}\rfloor\log\left[1+\lambda_1^2\frac{N^{1-\alpha}}{n^{2-\alpha}}\right]+o(1)\\
&\le \frac{1}{2}\lfloor\frac{2}{1-\alpha}\rfloor\log\left[(1+\lambda_1^2)\frac{N^{1-\alpha}}{n^{2-\alpha}}\right]+o(1)\\
&\le \frac{1}{1-\alpha}\log(1+\lambda_1^2)+o(1)+\log(N/n^2)\left\{\begin{array}{lll}-\frac{\alpha}{1-\alpha}\log(n) & \text{if} & \alpha\ge 1/3\\ -\alpha\log(N/n) & \text{if} & \alpha<1/3\end{array}\right. ,
\end{aligned}$$

where in the last line, we have used the identity $\lfloor 2/(1-\alpha)\rfloor=1$ for $\alpha<1/3$.

And we also have

$$\log\left(\frac{k_{\min}N}{n^2}\right)\ge\log\left(N/n^2\right)\ , \tag{73}$$

so that

$$\Delta\frac{k_{\min}-1}{2}-\log\left(\frac{k_{\min}N}{n^2}\right)\le\frac{1}{1-\alpha}\log(1+\lambda_1^2)-\left\{\begin{array}{lll}-\frac{\alpha}{1-\alpha}\log(n) & \text{if} & \alpha\ge 1/3\\ -\alpha\log(N/n) & \text{if} & \alpha<1/3\end{array}\right. ,$$

which goes to $-\infty$ since $\lambda_1=O(1)$ and $\alpha\log(N/n)=\lambda_0\to\infty$. Hence, we have $A_2=o(1)$.

- It remains to prove that $A_3=o(1)$. If we assume that $p_1\le 2p_0$, then $\Delta_k\le\Delta\le p_0(1+o(1))$ and we can prove that $A_3=o(1)$ arguing as for A_2 above:

$$\begin{aligned}
A_3 &\le \sum_{k=k_{\min}+1}^{n}\exp\left[k\left(\Delta\frac{k-1}{2}-\log\left(\frac{k}{n\rho}\right)+1\right)\right]\\
&\le \sum_{k=k_{\min}+1}^{n}\exp\left[k\left(1+o(1)+\max_{\ell\in\{k_{\min}+1,n\}}\left\{\Delta\frac{\ell-1}{2}-\log\left(\frac{\ell N}{n^2}\right)\right\}\right)\right]\\
&\le \sum_{k=k_{\min}+1}^{n}\exp\left[k\left(1+o(1)+\Delta\frac{n}{2}-\log\left(\frac{k_{\min}N}{n^2}\right)\right)\right]\ .
\end{aligned}$$

On one hand, we have $\Delta n \prec np_0 = (n/N)^{1-\alpha} = o(1)$. On the other hand, $\log(k_{\min} N/n^2) \to \infty$. Indeed, when $N^{1-\alpha} \leq n^{2-\alpha}$, we have (72); and when $N^{1-\alpha} > n^{2-\alpha}$, then $N/n^2 > n^{\alpha/(1-\alpha)} \to \infty$ and we use (73). We conclude that $A_3 = o(1)$ when $p_1 \leq 2p_0$. In the following, we suppose that $p_1 \geq 2p_0$. Leaving w_k unspecified, so we can use the same arguments later, we have

$$
\begin{aligned}
A_3 &= \mathbb{E}_0\left[\mathbb{1}_{\{k_0+1\leq K\leq k_{\min}\}} \exp\left(2\theta W_{S_1\cap S_2} - 2\Lambda(\theta)K^{(2)}\right)\mathbb{1}_{\{W_{S_1\cap S_2}\leq w_K\}}\right] \\
&= \sum_{k=k_{\min}+1}^{n}\sum_{i=1}^{w_k} \mathbb{P}_0\left[|S_1\cap S_2| = k, W_{S_1\cap S_2} = i\right] \exp\left[2i\theta - 2k^{(2)}\Lambda(\theta)\right] \\
&\leq \sum_{k=k_{\min}+1}^{n}\sum_{i=1}^{w_k} \binom{n}{k}\rho^k \binom{k^{(2)}}{i} p_0^i (1-p_0)^{k^{(2)}-i} \exp\left[2i\log\left(\frac{p_1}{p_0}\right) + 2(k^{(2)}-i)\log\left(\frac{1-p_1}{1-p_0}\right)\right] \\
&:= \sum_{k=k_{\min}+1}^{n}\sum_{i=1}^{w_k} B_{i,k} \ .
\end{aligned}
$$

Furthermore, since $0 < 1-p_0 < 1$ and $1-p_1 < 1-p_0$, we have

$$
B_{i,k} \leq \binom{n}{k}\rho^k \binom{k^{(2)}}{i} p_0^i (p_1/p_0)^{2i} \leq e^{o(k)}\left(\frac{en^2}{kN}\right)^k \left(\frac{ep_1^2 k^{(2)}}{p_0 i}\right)^i \ , \tag{74}
$$

using the standard bound $\binom{n}{k} \leq (en/k)^k$.

We now specify the calculations when $w_k = k$. Considering the sums over $i = 1, \ldots k/2$ and over $i = k/2+1, \ldots k$ separately, we get

$$
\begin{aligned}
\sum_{i=1}^{k} B_{i,k} &\leq e^{o(k)}\left(\frac{en^2}{kN}\right)^k \left[\sum_{i=1}^{\lfloor k/2\rfloor}\left(\frac{ep_1^2 k^{(2)}}{p_0}\right)^i + \sum_{\lfloor k/2\rfloor+1}^{k}\left(\frac{ep_1^2 k^{(2)}}{p_0 k/2}\right)^i\right] \\
&\leq e^{o(k)}\left(\frac{en^2}{kN}\right)^k k\left[1 + \left(\frac{ep_1^2 k^{(2)}}{p_0}\right)^{k/2} + \left(\frac{ep_1^2 k^{(2)}}{p_0 k/2}\right)^k\right] \\
&\prec e^{o(k)}\left[\left(\frac{en^2}{kN}\right)^k + \left(\frac{e^{3/2} n^2 p_1}{N\sqrt{2p_0}}\right)^k + \left(\frac{e^2 n^2 p_1^2}{Np_0}\right)^k\right] \ .
\end{aligned}
$$

First, $\frac{en^2}{kN} \leq \frac{en^2}{k_0 N} = o(1)$ by definition of k_0. Next, $\frac{n^2 p_1}{N\sqrt{p_0}} \leq \frac{2(p_1-p_0)}{\sqrt{p_0}}\frac{n^2}{N} = 2\sqrt{\zeta} \to 0$, by the fact that $p_1 \geq 2p_0$. Finally, $n^2 p_1^2/(Np_0) = \lambda_1^2/\lambda_0 \to 0$ since $\lambda_0 \to \infty$ and $\lambda_1 = O(1)$. Hence, we conclude that

$$
\sum_{k=k_{\min}+1}^{n}\sum_{i=1}^{k} B_{i,k} = o(1) \ . \tag{75}
$$

This immediately implies that $A_3 = o(1)$.

This concludes the proof of Theorem 4 under (62).

Proof of Lemma 20. Let us consider the event

$$
\Gamma'_S := \{\text{no connected component of } \mathcal{G}_S \text{ has more than one loop}\}
$$

Under Γ'_S, a connected component of $\mathcal{G}_S$ has at most as many edges as vertices. Consequently, $\Gamma_S \subset \Gamma'_S$ and we only need to prove that $\mathbb{P}_S(\Gamma'_S) = 1 + o(1)$. Since $\limsup \lambda_1 < 1$ and $\mathbb{P}_S(\Gamma'_S)$ is nondecreasing in λ_1, we may assume that λ_1 is fixed in $(0, 1)$.

As a warmup for what follows, we note that the number $\mathbf{L}_k$ of loops of size k in $\mathcal{G}_S$ satisfies

$$\mathbb{E}_S[\mathbf{L}_k] = p_1^k \frac{n!}{(n-k)!2k} \leq \frac{\lambda_1^k}{2k} \ ,$$

since there are $n!/[(n-k)!2k]$ potential loops of size k. Now, if a connected component contains (at least) two loops, there are two possibilities:

- The two loops have at least one edge in common. In that case, there is a loop (say of length k) with a chord (say of length $s < k$). Let $\mathbf{L}'_{k,s}$ denote the number of such configurations, There are $n!/[(n-k)!2k]$ potential loops of size k. Given a loop of size k, there are less than $\binom{k}{2}$ starting and ending nodes possible for the chord. Once these two nodes are set, there remains less than $n!/(n-s+1)!$ possibilities for the other nodes on the chord. Thus, we have

  $$\mathbb{E}_S[\mathbf{L}'_{k,s}] \leq p_1^{k+s} \frac{n!}{(n-k)!2k} \binom{k}{2} \frac{n!}{(n-s+1)!} \leq \left(\frac{\lambda_1}{n}\right)^{k+s} kn^{k+s-1} \leq \lambda_1^{k+s} \frac{k}{n} \ .$$

  Summing this inequality over s and k, we control the expected number of loops with a chord:

  $$\sum_{k=3}^{\infty} \sum_{s=1}^{k-1} \mathbb{E}[\mathbf{L}'_{k,s}] \leq \frac{1}{n} \sum_{k=3}^{\infty} \frac{k\lambda_1^{k+1}}{1-\lambda_1} \asymp \frac{1}{n} = o(1) \ ,$$

  since $\limsup \lambda_1 < 1$. Hence, this event occurs with probability going to 0.

- The two loops have no edge in common. Since there are in the same connected component, there is a path that goes from a vertex in the first loop to a vertex in the second loop. Let us note $\mathbf{L}'_{k_1,k_2,s}$ the number of loops of size k_1 and k_2 that do not share an edge and are connected by a path of length s. Observe that there are less $\frac{n!}{(n-k_1)!2k_1}$ possible configurations for the first loop, less than $\frac{n!}{(n-k_2)!2k_2}$ possible configurations for the second loop, and less than $k_1 k_2\, n!/(n-s+1)!$ possibilities for the chord. Thus, we get

  $$\begin{aligned} \mathbb{E}\left[\mathbf{L}'_{k_1,k_2,s}\right] &\leq p_1^{k_1+k_2+s} \frac{n!}{(n-k_1)!2k_1} \frac{n!}{(n-k_2)!2k_2} k_1 k_2 \frac{n!}{(n-s+1)!} \\ &\leq \left(\frac{\lambda_1}{n}\right)^{k_1+k_2+s} n^{k_1+k_2+s-1} = \frac{\lambda_1^{k_1+k_2+s}}{n} \ , \end{aligned}$$

  so that the expected number of such configurations is bounded as follows

  $$\sum_{k_1 \geq 3} \sum_{k_2 \geq 3} \sum_{s \geq 1} \mathbb{E}\left[\mathbf{L}'_{k_1,k_2,s}\right] \leq \frac{1}{n} \sum_{k_1 \geq 3} \sum_{k_2 \geq 3} \sum_{s \geq 1} \lambda_1^{k_1+k_2+s} \asymp \frac{1}{n} = o(1) \ .$$

  Hence, this second event occurs with probability going two zero

All in all, we have proved that $\mathbb{P}_S(\Gamma'_S) = 1 + o(1)$, implying that $\mathbb{P}_S(\Gamma_S) = 1 + o(1)$. $\square$

4.1.4 Proof of Theorem 4 under (63)

We follow the arguments laid out for the case (62). We define Γ_S in the same way, except that $w_k := \left[k\frac{(1-c)^{1/2}}{1-\alpha}\right]$, where c is a positive constant (to be chosen small later) such that $c < \alpha$ and, eventually,

$$\sup_{n/u_N < k \le n} \frac{1}{k}\,\mathbb{E}_S[W^*_{k,S}] \le \frac{1-2c}{1-\alpha}\ . \tag{76}$$

Lemma 21. *For any $k > k_{\min}$ and any subset S of size n, we have $\mathbb{P}_S[\Gamma_S] = 1 + o(1)$.*

For the second moment, we proceed exactly as in the case (62), and we start from (75). In fact, when $w_k \le k$, the proof is complete. So we assume that c is small enough that $w_k > k$, and bound the sum over $k+1 \le i \le w_k$. For $i > k$, we use the bound (74), together with the fact that $\lambda_0 = (N/n)^\alpha$ and $k < i$, to derive

$$\begin{aligned}
B_{i,k} &\le e^{o(k)}\left(\frac{en^2}{kN}\right)^k\left(\frac{ep_1^2k^{(2)}}{p_0 i}\right)^i \\
&\le e^{o(k)}\left(\frac{en^2}{kN}\right)^k\left(\frac{N^{1-\alpha}k}{n^{2-\alpha}}\frac{\lambda_1^2 e}{2}\right)^i \\
&= e^{o(k)+k}\left(\frac{n}{N}\right)^{k-i(1-\alpha)}\left(\frac{\lambda_1^2 e}{2}\right)^i\left(\frac{n}{k}\right)^{k-i} \\
&\le e^{o(k)+k}\left(\frac{n}{N}\right)^{k-i(1-\alpha)}\left(\frac{\lambda_1^2 e}{2}\right)^i\ .
\end{aligned}$$

This allows us to control the sum

$$\begin{aligned}
\sum_{i=k+1}^{w_k} B_{i,k} &\le w_k\, e^{o(k)+k}\left(\frac{n}{N}\right)^{k-(1-\alpha)w_k}\left(\frac{\lambda_1^2 e}{2}\vee 1\right)^{w_k} \\
&\prec k e^{o(k)+k}\left(\frac{n}{N}\right)^{k(1-(1-c)^{1/2})}\left(\frac{\lambda_1^2 e}{2}\vee 1\right)^{k\frac{(1-c)^{1/2}}{1-\alpha}} \\
&= \exp\left[O(k) - k(1-(1-c)^{1/2})\log(N/n)\right]\ ,
\end{aligned}$$

where in the second line we used the fact that $w_k = O(k)$ since $\limsup a < 1$, and in the third line we used the fact that $\lambda_1 = O(1)$. Thus,

$$\sum_{k=k_{\min}+1}^{n}\sum_{i=k+1}^{w_k} B_{i,k} = o(1)\ ,$$

which together with (75) allows us to conclude that $A_3 = o(1)$.

This concludes the proof of Theorem 4 under (63).

Proof of Lemma 21. Recall that $u_N = \log\log(N/n)$. First we consider integers k satisfying $k_{\min} + 1 \le k < n/u_N$. Define $\omega'_k = k(1-c)^{-1/2}\left(\frac{\lambda_1}{2}\vee 1\right)$ and $q'_k = \omega'_k/k^{(2)}$. Applying a union bound and Chernoff's bound for the binomial distribution, we derive that

$$\begin{aligned}
\mathbb{P}_S\left[W^*_{k,S} \ge \omega'_k\right] &\le \binom{n}{k}\mathbb{P}[\mathrm{Bin}(k^{(2)}, p_1) \ge \omega'_k] \\
&\le \exp\left[k\left\{\log\left(\frac{ne}{k}\right) - \frac{k-1}{2}H_{p_1}(q'_k)\right\}\right]\ .
\end{aligned}$$

Since $k/n \leq 1/u_N = o(1)$, and since λ_1 is bounded, we have $q'_k/p_1 \to \infty$, so that

$$\begin{aligned}\frac{k-1}{2} H_{p_1}(q'_k) &\sim \frac{k-1}{2} q'_k \log\left(\frac{q'_k}{p_1}\right) \\ &= (1-c)^{-1/2}\left[\frac{\lambda_1}{2} \vee 1\right]\left[\log\left(\frac{n}{k-1}\right) + \log\left\{(1-c)^{-1/2}\left(1 \vee \frac{2}{\lambda_1}\right)\right\}\right] \\ &\geq (1+o(1))(1-c)^{-1/2} \log\left(\frac{n}{k}\right) ,\end{aligned}$$

and therefore, since $c \in (0,1)$ is fixed,

$$\log\left(\frac{ne}{k}\right) - \frac{k-1}{2} H_{p_1}(q'_k) \leq 1 + \left[1 - (1+o(1))(1-c)^{-1/2}\right] \log(u_N) \to -\infty .$$

We conclude that

$$\sum_{k=k_{\min}+1}^{n/u_N} \mathbb{P}_S\left[W^*_{k,S} \geq \omega'_k\right] = o(1) .$$

Let us now prove that $\omega'_k \leq w_k$. Indeed, this inequality holds if, and only if, $\lambda_1 \leq 2(1-c)/(1-\alpha)$ and $c \leq \alpha$. The second inequality is by definition of c, while the first inequality is ensured by (76) since

$$\frac{\lambda_1}{2}\frac{n-1}{n} = \mathbb{E}_S[W^*_{n,S}/n] \leq \sup_{k \leq n} \mathbb{E}_S[W^*_{k,S}/k] \leq (1-2c)/(1-\alpha) .$$

Let us turn to integers k satisfying $k \geq n/u_N$. Let $c_0 = (1-c)^{-1/2} - 1$ and $t = c_0 \, \mathbb{E}_S[W^*_{k,S}]$. By taking any fixed subset $T \subset S$ of size $|T| = k$, we derive

$$\mathbb{E}_S[W^*_{k,S}] \geq \mathbb{E}_S[W_T] = p_1 k^{(2)} \geq \frac{\lambda_1}{n}(n/u_N)^{(2)} \asymp \frac{n}{u_N^2} \to \infty , \tag{77}$$

so that t satisfies the condition of Lemma 7 eventually. Using that lemma, we derive that

$$\mathbb{P}_S\left[W^*_{k,S} \geq \mathbb{E}_S[W^*_{k,S}](1-c)^{-1/2}\right] \leq \exp\left[-\mathbb{E}_S[W^*_{k,S}]\frac{\log(2)}{4} c_0 \left[1 \wedge \frac{c_0}{8}\right]\right]$$

By Condition (76), $w_k \geq \mathbb{E}_S[W^*_{k,S}](1-c)^{-1/2}$. Hence, there exists a positive constant κ, such that

$$\begin{aligned}\sum_{k=n/u_N}^{n} \mathbb{P}_S\left[W^*_{k,S} \geq w_k\right] &\leq \sum_{k=n/u_N}^{n} \exp\left[-\kappa\, \mathbb{E}_S[W^*_{k,S}]\right] \\ &\leq n \exp\left[-\kappa\, \mathbb{E}_S\left[W^*_{\frac{n}{u_N},S}\right]\right]\end{aligned}$$

Because of (77) and the fact that $\log(N) = o(n)$, we have

$$\mathbb{E}_S\left[W^*_{\frac{n}{u_N},S}\right] \succ \frac{n}{\log^2(n)} ,$$

and therefore the sum above goes to 0. $\square$

4.2 No test is asymptotically powerful

When λ_0 is bounded away from 0 and infinity, the triangle test has some non-negligible power as long as λ_1 is bounded away from 0 (see Section 3.4). This motivates us to obtain sufficient conditions under no test is asymptotically powerful.

Our method is also based on bounding the first two moments of a truncated likelihood ratio $\tilde{L}$. Indeed, it is enough to show that $\liminf \mathbb{E}_0 \tilde{L} > 0$ and $\liminf \mathbb{E}_0[\tilde{L}^2] < \infty$. This comes from the following result.

Lemma 22. *Let $\mathbb{P}_0$ and $\mathbb{P}_1$ be two probability distributions on the same probability space, with densities f_0 and f_1 with respect to some dominating measure. Let Γ be any event and define the truncated likelihood ratio $\tilde{L} = L\,\mathbb{1}_\Gamma$, where $L = f_1/f_0$ is the likelihood ratio for testing $\mathbb{P}_0$ versus $\mathbb{P}_1$. Then any test for $\mathbb{P}_0$ versus $\mathbb{P}_1$ has risk at least*

$$\frac{4}{27}\frac{(\mathbb{E}_0\,\tilde{L})^3}{\mathbb{E}_0[\tilde{L}^2]}\ ,$$

where $\mathbb{E}_0$ denotes the expectation under $\mathbb{P}_0$, and by convention $0/0 = 0$.

Proof. Assume $\mathbb{E}_0\,\tilde{L} \neq 0$, for otherwise the result is immediate. The risk of the likelihood ratio test $\{L > 1\}$ — which is the test that optimizes the risk — is equal to

$$B := 1 - \frac{1}{2}\,\mathbb{E}_0\,|L-1| = 1 - \mathbb{E}_0(1-L)_+ \geq 1 - \mathbb{E}_0(1-\tilde{L})_+\ ,$$

since $\tilde{L} \leq L$. For any $t \in (0,1)$, we have

$$\mathbb{E}_0(1-\tilde{L})_+ \leq (1-t)\,\mathbb{P}_0(\tilde{L} > t) + \mathbb{P}_0(\tilde{L} \leq t) = 1 - t\,\mathbb{P}_0(\tilde{L} > t)\ .$$

Moreover, using the Cauchy-Schwarz inequality, we have for any $t > 0$

$$\begin{aligned}
\mathbb{E}_0\,\tilde{L} &= \mathbb{E}_0[\tilde{L}\,\mathbb{1}_{\{\tilde{L}\leq t\}}] + \mathbb{E}_0[\tilde{L}\,\mathbb{1}_{\{\tilde{L}>t\}}] \\
&\leq t + \sqrt{\mathbb{E}_0[\tilde{L}^2]\ \ \mathbb{P}_0(\tilde{L} > t)}\ ,
\end{aligned}$$

so that, taking $t < \mathbb{E}_0\,\tilde{L}$, we have

$$\mathbb{P}_0(\tilde{L} > t) \geq \frac{(\mathbb{E}_0\,\tilde{L} - t)^2}{\mathbb{E}_0\,\tilde{L}^2}\ .$$

We conclude that

$$B \geq t\,\mathbb{P}_0(\tilde{L} > t) \geq t\ \frac{(\mathbb{E}_0\,\tilde{L} - t)^2}{\mathbb{E}_0\,\tilde{L}^2}\ ,$$

and optimizing this over $0 < t < \mathbb{E}_0\,\tilde{L}$ yields the result. □

Since we only need to focus on the case where λ_0 is bounded from 0 and infinity, and where λ_1 is bounded from 0 (because the other cases are covered by Theorem 4), we may assume they are fixed without loss of generality. In that case $\zeta \to 0$ is equivalent to $n^2/N \to 0$, which is what we assume in the following.

Theorem 5. *Write $n = N^\kappa$ with $0 < \kappa < 1/2$, and assume that λ_0 and λ_1 are both fixed. No test is asymptotically powerful in all the following situations:*

$$\lambda_1 < 1, \quad \lambda_1^2 e \leq \lambda_0 \ ; \tag{78}$$

$$\lambda_1 < 1, \quad \lambda_1^2 e > \lambda_0, \quad \frac{1-2\kappa}{\kappa} \frac{I_{\lambda_1}}{\log\left(\frac{e\lambda_1^2}{\lambda_0}\right)} > 1 \ . \tag{79}$$

Proof of Theorem 5. We use the same truncation as in Section 4.1.2, and still denote the resulting truncated likelihood by $\tilde{L}$.

For the first moment, by symmetry,

$$\mathbb{E}_0[\tilde{L}] = \mathbb{P}_S[\Gamma_S] = \mathbb{P}_S[\mathcal{G}_S \text{ is a forest}, \ |\mathcal{C}_{\max,S}| \leq f_n] \ .$$

We already saw that $\mathbb{P}_S[|\mathcal{C}_{\max,S}| \leq f_n] \to 1$ (Van der Hofstad, 2012, Th. 4.4). Consequently,

$$\mathbb{E}_0[\tilde{L}] = \mathbb{P}_S[\mathcal{G}_S \text{ is a forest}] + o(1) \ .$$

Of course, $\mathcal{G}_S$ is a forest if, and only if, it has no cycles. By Takács (1988), the number of cycles in $\mathcal{G}_S$ converges weakly to a Poisson distribution with mean

$$a(\lambda_1) = \frac{1}{2}\log\left(\frac{1}{1-\lambda_1}\right) - \frac{\lambda_1}{2} - \frac{\lambda_1^2}{4} \ ,$$

when $\lambda_1 < 1$ is fixed. As a consequence, $\mathbb{E}_0[\tilde{L}] = \exp[-a(\lambda_1)] + o(1)$, which remains bounded away from zero.

For the second moment, we start from (70):

$$\mathbb{E}_0\left[\tilde{L}^2\right] - 1 \prec \sum_{j=1}^{\infty} \left(\frac{n^2 e}{N}\left[1 \vee \frac{\lambda_1^2 e}{\lambda_0}\right]^{f_n}\right)^j \ ,$$

with $f_n = (1+c)I_{\lambda_1}^{-1}\log n$ and c is a small positive constant. Under (78), we have $\lambda_1^2 e \leq \lambda_0$ and the RHS is $O(n^2/N) = o(1)$. Under (79), we have $\lambda_1^2 e > \lambda_0$, and the RHS is, as before, equal to (71). Here we have

$$A_n = \left[1 - 2\kappa - (1+c)\frac{\kappa}{I_{\lambda_1}}\log\left(\frac{\lambda_1^2 e}{\lambda_0}\right)\right]\log N \to \infty \ ,$$

when (79) is satisfied and c is small enough. Hence, in any case, we found that $\mathbb{E}_0\left[\tilde{L}^2\right] \leq 1 + o(1)$. □

5 Some tests running in polynomial time

There is a surge of interest in decision theory with computational constraints, the main question being, what can be achieved with tests that can be computed in polynomial-time. An example is the *planted clique problem* mentioned in the Introduction, where it is conjectured that there are no polynomial-time algorithms that can detect with high confidence the presence of a clique of size k added to $\mathcal{G} \sim \mathbb{G}(N, 1/2)$.

To focus the discussion, assume that the total degree test (which runs in $O(N^2)$ time at most) is *not* asymptotically powerful. In fact, to be even more specific, assume that $\limsup \lambda_1 < \infty$ and

that $n^2/N = o(1)$. In this setting, the largest connected component test — which can be computed in $O(|\mathcal{E}|+|\mathcal{V}|) = O(N^2)$ time — is powerful when $\liminf \lambda_0 < 1$ and λ_1 is sufficiently large. And the triangle test — which can be computed in $O(N^{2.82})$ time (Alon et al., 1997) — is *not* asymptotically powerful unless $\lambda_0 \to 0$. Those are the only tests that we studied here which we can compute in polynomial time. The k-tree test is powerful when λ_0 is sufficient small and λ_1 is sufficiently large, but we do not know how to compute it in polynomial-time. And, similarly, the broad scan test is asymptotically optimal when $\lambda_0 \geq e$, but we do not know of any polynomial-time algorithm for computing it.

In this section, we discuss a few other tests that can be computed in polynomial time. Here is a brief summary of the best detection bounds and the corresponding tests. In the poissonian asymptotic where λ_0 and λ_1 are both fixed, a polynomial time approximation of the number of k-cycles (with k large) is powerful as soon as $\lambda_1 > \sqrt{\lambda_0} \vee 1$. In the asymptotic $\lambda_0 \to \infty$ with $\log(N) \prec \lambda_0 \prec n\log(N)$, the test based on the second eigenvalue of $\mathbf{W}$ is powerful for $\lambda_1 \gg \sqrt{\lambda_0}$. For $n\log(N) \prec \lambda_0 \prec N/n$, a convex relaxation of the n-sparse eigenvalue problem is powerful for $\lambda_1 \gg \sqrt{n\log(N)} \bigvee (\lambda_0 \log(N))^{1/4} \bigvee \sqrt{\lambda_0}\left(\frac{n^2\log(N)}{N}\right)^{1/4}$.

5.1 The maximal degree test

The test based on the maximum degree was examined in (Arias-Castro and Verzelen, 2012) under the assumption that $\lambda_0 \gg \log N$, and was shown to be asymptotically powerful when $\lambda_1 \gg \sqrt{\lambda_0 \log N}$. We now concentrate on the case where $\lambda_0 \prec \log N$.

Recall the function $h(x) = x\log x - x + 1$ that appears in Lemma 2. Below, h^{-1} refers to the inverse of $h : [1,\infty) \to [0,\infty)$, where it is strictly increasing. We will also use the fact that

$$h(ab) = ah(b) + bh(a) + (a-1)(b-1) \ , \quad \forall a, b > 0 \ . \tag{80}$$

Proposition 5. *Assume that $n = N^\kappa$ with $\kappa \in (0, 1/2)$ fixed, and that $\lambda_0 = \gamma_0 \log N$ and $\lambda_1 = \gamma_1 \log N$, where $\gamma_0 \vee \gamma_1 = O(1)$ and $-\log(\gamma_0 \wedge \gamma_1) = o(\log N)$. Then the maximum degree test is asymptotically powerful if*

$$\liminf \xi > 1 \ , \quad \xi := \big(1 + \frac{\gamma_1}{\gamma_0}\kappa\big)\frac{h^{-1}\big(\frac{\kappa}{\gamma_0+\gamma_1\kappa}\big)}{h^{-1}\big(\frac{1}{\gamma_0}\big)} \ ,$$

and powerless if $\limsup \xi < 1$. *In particular, when $\gamma_0 \to 0$, the test is powerful if*

$$\liminf \xi' > 1 \ , \quad \xi' := \frac{\kappa \log(1/\gamma_0)}{\log(1/\gamma_1)} \ ,$$

and powerless if $\limsup \xi' < 1$.

The proposition implies that the test is powerless when λ_0 and λ_1 are both fixed. The regime where $\lambda_0 = O(N^{-a_0})$ for some $a_0 > 0$ fixed is excluded. However, in this very sparse regime, the maximum degree converges to a constant — except for special circumstances, see (Bollobás, 2001, Th. 3.2) — and the statement of Proposition 5 would have to be modified with careful rounding.

Proof of Proposition 5. Let $\Delta(\mathcal{H})$ denote the maximum degree of a graph $\mathcal{H}$. Below Δ is a shorthand for $\Delta(\mathcal{G})$. We first prove that, in probability under $\mathbb{P}_0$,

$$\frac{\Delta}{\gamma_0 h^{-1}(1/\gamma_0)\log N} \to 1 \ . \tag{81}$$

Let $p_0 = \lambda_0/N$ and $\eta = h^{-1}(\log(N)/\lambda_0)$.

- *Upper bound.* First, let $k = \lceil (1+\varepsilon)\eta\lambda_0 \rceil$. The expected number of nodes with degree at least k is equal to

$$\begin{aligned} N\,\mathbb{P}\big[\mathrm{Bin}(N-1,p_0) \geq k\big] &\leq N\exp[-NH_{p_0}(k/N)] \\ &\leq \exp\big[\log N - \lambda_0 h((1+\varepsilon)\eta) + O(\lambda_0^2\eta^2/N)\big]\ , \end{aligned}$$

using Lemma 1 and then Lemma 2, with the fact that $H_{p_0}(k/N) \geq H_{p_0}((1+\varepsilon)\eta p_0)$. Note that $\lambda_0^2\eta^2/N = O(\log(N)/N) = o(1)$ in our context. We then use (80), to get

$$\log N - \lambda_0 h((1+\varepsilon)\eta) \leq -\varepsilon \log N \to -\infty\ .$$

We conclude with Markov inequality that $\mathbb{P}_0(\Delta \geq k) \to 0$.

- *Lower bound.* For the lower bound, redefine $k = \lfloor (1-\varepsilon)\eta\lambda_0 \rfloor$. We could use (Bollobás, 2001, Th. 3.2), but we will derive the result 'by hand', because the same argument is used below. Let D_k denote the number of nodes with degree exactly k, so that $D_k = \sum_i \mathbb{1}_{\{W_i=k\}}$, where W_i is the degree of node i. Note that $W_i \sim \mathrm{Bin}(N-1,p_0)$ under $\mathbb{P}_0$. Using Stirling's formula, we have

$$\begin{aligned} \mathbb{E}_0[D_k] &= N\,\mathbb{P}[\mathrm{Bin}(N-1,p_0) = k] \\ &\asymp \exp\big[\log N - \frac{1}{2}\log k - (N-1)H_{p_0}(k/(N-1))\big] \\ &\geq \exp\big[\log N - \frac{1}{2}\log k - \lambda_0 h((1-\varepsilon)\eta) + O(\lambda_0^2\eta^2/m)\big]\ , \end{aligned}$$

applying Lemma 2 in the third line. We note that $\lambda_0^2\eta^2/N = o(1)$ as before, and

$$k \leq \eta\lambda_0 = O\big(\frac{\log N}{\log(\log(N)/\lambda_0))}\big) = O(\log N)\ ,$$

so that $\log k = O(\log\log N)$. Applying (80), we then get

$$\log N - \lambda_0 h((1-\varepsilon)\eta) \geq \big[\varepsilon - \frac{\eta}{h(\eta)} h(1-\varepsilon)\big] \log N\ .$$

Letting $c = \liminf \log(N)/\lambda_0 > 0$, we have $\eta/h(\eta) \leq h^{-1}(c)/c$, and $\varepsilon - \frac{\eta}{h(\eta)}h(1-\varepsilon) \geq \varepsilon - \frac{h^{-1}(c)}{c}h(1-\varepsilon) > 0$ for $\varepsilon > 0$ small enough. We therefore have that $\mathbb{E}_0[D_k] \to \infty$. We now turn to the variance. Let $q_k = \mathbb{P}[\mathrm{Bin}(N-2,p_0) = k]$. For $i,j \in \mathcal{V}$ distinct, conditioning on W_{ij}, we get

$$\mathbb{P}_0[W_i = k, W_j = k] = (1-p_0)q_k^2 + p_0 q_{k-1}^2\ ,$$

while

$$\mathbb{P}_0[W_i = k] = (1-p_0)q_k + p_0 q_{k-1}\ ,$$

obtaining

$$\mathrm{Cov}_0(\mathbb{1}_{\{W_i=k\}}, \mathbb{1}_{\{W_j=k\}}) = p_0(1-p_0)(q_k - q_{k-1})^2\ .$$

Eventually,

$$\mathrm{Var}_0[D_k] \leq \mathbb{E}_0[D_k] + N(N-1)p_0(1-p_0)(q_k - q_{k-1})^2\ ,$$

so that

$$\begin{aligned}
\frac{\mathrm{Var}_0[D_k]}{(\mathbb{E}_0[D_k])^2} &\leq \frac{1}{\mathbb{E}_0[D_k]} + \frac{N^2 p_0 (q_k - q_{k-1})^2}{N^2((1-p_0)q_k + p_0 q_{k-1})^2} \\
&= \frac{1}{\mathbb{E}_0[D_k^S]} + O(p_0)\big(1 + \frac{q_{k-1}^2}{q_k^2}\big) \\
&= \frac{1}{\mathbb{E}_0[D_k^S]} + O(p_0) + O(\frac{k^2}{p_0 N^2}) \to 0 \ ,
\end{aligned}$$

since $p_0 \frac{q_{k-1}^2}{q_k^2} \sim \frac{k^2}{p_0 N^2} \prec \frac{(\log N)^2}{\lambda_0 N} \to 0$. Hence, by Chebyshev inequality, $D_k \to \infty$ under $\mathbb{P}_0$, implying that $\mathbb{P}_0(\Delta \geq k) \to 1$.

The proof of (81) is now complete.

Henceforth, we work under the alternative $\mathbb{P}_S$. Define $\tilde{\Delta}_S := \max_{i \in S} W_i$. We have $W_i = W_i^S + W_i^{S^c}$, where $W_i^T = \sum_{j \in T} W_{ij}$ for $T \subset \mathcal{V}$. (Recall that $W_{ii} = 0$.) Our arguments are parallel to those we used under $\mathbb{P}_0$, the only difficulty being that W_i is not binomial anymore. Indeed, $W_i^S \sim \mathrm{Bin}(n-1, p_1)$ and $W_i^{S^c} \sim \mathrm{Bin}(N-n, p_0)$ are independent. Nevertheless, the resulting Poisson binomial distribution is close to a binomial distribution.

Lemma 23. *Suppose $X_1 \sim Bin(m_1, q_1)$ and $X_2 \sim Bin(m_2, q_2)$ are independent. Let $m = m_1 + m_2$ and $q = \frac{m_1 q_1 + m_2 q_2}{m_1 + m_2}$. Then, for $k \geq mq$,*

$$\mathbb{P}[X_1 + X_2 \geq k] \leq \mathbb{P}[Bin(m,q) \geq k] \ .$$

If $q_1 \vee q_2 < 1/2$ and $k < (m_1 \wedge m_2)/2$, we also have

$$\exp\left[-qk - \frac{(mq)^2 + k^2}{m_1 \wedge m_2}\right] \leq \frac{\mathbb{P}[X_1 + X_2 = k]}{\mathbb{P}[Bin(m,q) = k]} \leq \exp\left[(q_1 \vee q_2)k + \frac{(mq)^2 + k^2}{m}\right] \ .$$

Equipped with Lemma 23, we proceed with controlling $\tilde{\Delta}_S$ under $\mathbb{P}_S$ using the same arguments. In fact, we prove that, under $\mathbb{P}_S$,

$$\frac{\tilde{\Delta}_S}{\Big[(\gamma_1 \kappa + \gamma_0) h^{-1}\big(\frac{\kappa}{\gamma_0 + \gamma_1 \kappa}\big)\Big] \log N} \to 1 \ . \tag{82}$$

- *Upper bound.* Let $\bar{p} = \frac{(n-1)p_1 + (N-n)p_0}{N-1}$ and $\bar{\lambda} = (N-1)\bar{p}$. For $\varepsilon > 0$ small, consider $k = (1+\varepsilon)\bar{\lambda} h^{-1}\big(\log(n)/\bar{\lambda}\big)$. As before, $k = O(\log N)$, and we have

$$\begin{aligned}
\mathbb{P}_S\big[\tilde{\Delta}_S \geq k\big] &\leq n\, \mathbb{P}\big[\mathrm{Bin}(N-1, \bar{p}) \geq k\big] \\
&\leq n \exp\big[-(N-1) H_{\bar{p}}(k/(N-1))\big] \\
&= \exp\big[\log n - \bar{\lambda}\, h(k/\bar{\lambda}) + O(k^2/N)\big] \\
&\leq \exp\big[-\varepsilon \log n + O((\log N)^2/N)\big] \to 0 \ ,
\end{aligned}$$

  The first line comes from the union bound and Lemma 23, the second line from Lemma 1, the third from Lemma 2, and the fourth from (80). Now, since $\bar{\lambda} \sim \lambda_1 + \lambda_0$, it is easy to see that

$$\bar{\lambda} h^{-1}\big(\log(n)/\bar{\lambda}\big) \sim (\lambda_1 + \lambda_0) h^{-1}\big(\log(n)/(\lambda_1 + \lambda_0)\big) \sim \Big[(\gamma_1 \kappa + \gamma_0) h^{-1}\big(\frac{\kappa}{\gamma_0 + \gamma_1 \kappa}\big)\Big] \log N \ .$$

We conclude that, in probability under $\mathbb{P}_S$,

$$\limsup \frac{\tilde{\Delta}_S}{\Big[(\gamma_1\kappa+\gamma_0)h^{-1}\big(\frac{\kappa}{\gamma_0+\gamma_1\kappa}\big)\Big]\log N} \le 1 \ .$$

- *Lower bound.* Let D_k^S denote the number of nodes in S with degree equal to k. Obviously, $\Delta \ge k$ if $D_k^S \ge 1$. Fix ε small, and redefine $k = \lfloor (1-\varepsilon)\bar{\lambda}h^{-1}\big(\log(n)/\bar{\lambda}\big)\rfloor$. We have

$$\begin{aligned}
\mathbb{E}_S[D_k^S] \ = \ n\,\mathbb{P}_S[W_i = k] \ &\ge \ n e^{-\bar{p}k - \frac{\bar{\lambda}^2+k^2}{n-1}}\,\mathbb{P}[\mathrm{Bin}(N-1,\bar{p}) = k] \\
&\succ \ \frac{n}{\sqrt{k}} \exp\big[-(N-1)H_{\bar{p}}(k/(N-1))\big] \\
&= \ \exp\Big[\log n - \frac{1}{2}\log k - \bar{\lambda}\,h(k/\bar{\lambda}) + O(k^2/N)\Big] \to \infty \ ,
\end{aligned}$$

where we used Lemma 23 in the first inequality; in the second the fact that $\bar{p}k + \frac{\bar{\lambda}^2+k^2}{n-1} = o(1)$ since $k = O(\log N)$, $\bar{\lambda} = O(\log N)$, $\bar{p} \le p_1 = O(\log(N)/n)$ and $n = N^\kappa$, as well as Stirling's inequality; we then applied Lemma 2. The divergence to infinity follows from the same arguments that we used before. We now bound the variance of D_k^S. Redefine $q_k = \mathbb{P}[X+Y = k]$ where $X \sim \mathrm{Bin}(n-2, p_1)$ and $Y \sim \mathrm{Bin}(N-n, p_0)$ are independent. Working exactly as we did before, we get

$$\frac{\mathrm{Var}_S[D_k^S]}{(\mathbb{E}_S[D_k^S])^2} \le \frac{1}{\mathbb{E}_S[D_k^S]} + p_1\big(1 + \frac{q_{k-1}^2}{q_k^2}\big) \ .$$

Let $\tilde{q}_k = \mathbb{P}[\mathrm{Bin}(N-2,\tilde{p}) = k]$, where $\tilde{p} := \frac{(n-2)p_1+(N-n)p_0}{N-2}$, and also $\tilde{\lambda} = (N-2)\tilde{p}$. By Lemma 23,

$$\tilde{q}_k\, e^{-\tilde{p}k - \frac{\tilde{\lambda}^2+k^2}{n-1}} \le q_k \le \tilde{q}_k\, e^{p_1 k + \frac{\tilde{\lambda}^2+k^2}{N-1}} \ ,$$

so that $q_k \sim \tilde{q}_k$, and therefore

$$p_1\frac{q_{k-1}^2}{q_k^2} \sim p_1\frac{\tilde{q}_{k-1}^2}{\tilde{q}_k^2} \sim p_1\frac{k^2}{\tilde{p}^2N^2} \prec \frac{\lambda_1}{\lambda_0^2}\frac{(\log N)^2}{n} = o(1) \ .$$

Hence, $N_k \to \infty$ under $\mathbb{P}_S$, by Chebyshev inequality. This implies that $\tilde{\Delta}_S \ge k$ with probability tending to one under $\mathbb{P}_S$. Then, arguing as for the upper bound, we conclude that under $\mathbb{P}_S$

$$\liminf \frac{\tilde{\Delta}_S}{\Big[(\gamma_1\kappa+\gamma_0)h^{-1}\big(\frac{\kappa}{\gamma_0+\gamma_1\kappa}\big)\Big]\log N} \ge 1 \ .$$

So we proved (82), and together with (81), we can conclude. $\square$

Proof of Lemma 23. The upper bound is a special case of (Hoeffding, 1956, Th. 4). Letting $\lambda_j = m_j q_j$ and $\lambda = mq$, the lower bound goes as follows:

$$\begin{aligned}
\mathbb{P}[X_1 + X_2 = k] \ &= \ \sum_{k_1+k_2=k} \mathbb{P}[\mathrm{Bin}(m_1,q_1) = k_1]\,\mathbb{P}[\mathrm{Bin}(m_2,q_2) = k_2] \\
&\ge \ \sum_{k_1+k_2=k} e^{-\frac{\lambda_1^2+k_1^2}{m_1}} e^{-\lambda_1}\frac{\lambda_1^{k_1}}{k_1!}\, e^{-\frac{\lambda_2^2+k_2^2}{m_2}} e^{-\lambda_2}\frac{\lambda_2^{k_2}}{k_2!} \\
&\ge \ e^{-\frac{\lambda^2+k^2}{m_1\wedge m_2}} e^{-\lambda}\frac{\lambda^k}{k!} \\
&\ge \ e^{-qk - \frac{(mq)^2+k^2}{m_1\wedge m_2}}\,\mathbb{P}[\mathrm{Bin}(m,q) = k] \ ,
\end{aligned}$$

using (Bollobás, 2001, Eq. 1.14) in the second line and (Bollobás, 2001, Eq. 1.13) in the fourth line. The upper bound is obtained similarly. □

5.2 A relaxation of the n-sparse eigenvalue problem

We also studied in (Arias-Castro and Verzelen, 2012) a test based on a convex relaxation of the n-sparse eigenvalue problem. Formally, for a positive semidefinite matrix $\mathbf{B} \in \mathbb{R}^{N\times N}$ and $1 \le n \le N$, define

$$\beta_n^{\max}(\mathbf{B}) = \max_{|S|=n} \|\mathbf{B}_S\| \ ,$$

where $\mathbf{B}_S$ denotes the principal submatrix of $\mathbf{B}$ indexed by $S \subset \{1,\dots,N\}$ and $\|\mathbf{B}\|$ the largest eigenvalue of $\mathbf{B}$. d'Aspremont et al. (2007) relaxed this to

$$\mathrm{SDP}_n(\mathbf{B}) = \max_{\mathbf{Z}} \mathrm{trace}(\mathbf{BZ}), \qquad \text{subject to } \mathbf{Z} \succeq 0,\ \mathrm{trace}(\mathbf{Z}) = 1,\ |\mathbf{Z}|_1 \le n \ ,$$

where the maximum is over positive semidefinite matrices $\mathbf{Z} = (Z_{st}) \in \mathbb{R}^{N\times N}$ and $|\mathbf{Z}|_1 = \sum_{s,t} |Z_{st}|$. We considered in (Arias-Castro and Verzelen, 2012) the relaxed scan test, which rejects for large values of $\mathrm{SDP}_n(\mathbf{W}^2)$. With the same method of proof that we used there, we can conclude that the test is asymptotically powerful when $n = N^\kappa$ with $0 < \kappa < 1/2$ fixed, $\lambda_0 \ll \frac{N}{n}$, and

$$\lambda_1 \gg \sqrt{n\log(N)} \bigvee (\lambda_0 \log(N))^{1/4} \bigvee \sqrt{\lambda_0}\left(\frac{n^2\log(N)}{N}\right)^{1/4} .$$

5.3 The number of k-cycles

Let C_k denote the number of simple k-cycles in the graph. To test against a stochastic block model alternative, Mossel et al. (2012) use the test based on C_k, with $k \to \infty$ slowly. Our arguments are also based on the first two moments of C_k, but controlling the variance is more involved.

Proposition 6. *The test based on C_k, with $k \to \infty$ and $k = O(\log N)^{1/4}$, is asymptotically powerful when λ_0 and λ_1 are fixed with $\lambda_1 > \sqrt{\lambda_0} \vee 1$.*

Although computing the number of simple k-cycles seems difficult, Alon and Gutner (2010) provide an approximation that suffices for our purposes. They show that, for any $\delta \in (1,2)$, there is a (deterministic) function $\tilde{C}_k$ that can be computed in less than $g(k,\delta)N^3 \log N$ time on graphs with N nodes, such that

$$\delta^{-1} \le \frac{\tilde{C}_k(\mathcal{H})}{C_k(\mathcal{H})} \le \delta \ ,$$

for any fixed graph $\mathcal{H}$. In fact, $g(k) = e^{O(k\log\log k - k\log(1-\delta))}$.

Proposition 7. *Take $\delta = 1 + \exp[-\log^2(\log(N))]$ (for example). The test based on $\tilde{C}_k$, with $k \to \infty$ and $k = O(\log\log N)$ runs in polynomial time and is asymptotically powerful when λ_0 and λ_1 are fixed with $\lambda_1 > \sqrt{\lambda_0} \vee 1$.*

Proof of Proposition 6. We aim at applying (47). Straightforward calculations show that

$$\mathbb{E}_0\, C_k = \frac{N!}{(N-k)!2k} p_0^k \sim \frac{\lambda_0^k}{2k} \ , \tag{83}$$

since there are $\frac{N!}{(N-k)!2k}$ possible cycles with k edges, and $k^2 = o(N)$. For the variance, arguing as in (Bollobás, 2001, Eq. 4.2,4.4), which is based on bounding the number of pairs of cycles that share at least one node, we get

$$\begin{aligned}
\mathbb{E}_0[C_k(C_k-1)] &\leq \frac{N!p_0^{2k}}{(N-2k)!(2k)^2} + \sum_{\ell=k}^{2k-1} \binom{N}{\ell} \left[\binom{\ell}{k}\frac{k!}{2k}\right]^2 p_0^{\ell+\frac{1}{k}} \\
&\leq \left(\frac{N!p_0^k}{(N-k)!(2k)}\right)^2 + p_0^{1/k}\frac{1}{(2k)^2}\sum_{\ell=k}^{2k-1}\lambda_0^\ell \frac{\ell!}{[(\ell-k)!]^2} \\
&\leq \big(\mathbb{E}_0[C_k]\big)^2 + N^{-1/k}(\lambda_0\vee 1)^{2k}(2k)!\ .
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathrm{Var}_0[C_k] &= \mathbb{E}_0[C_k] + \mathbb{E}_0[C_k(C_k-1)] - \big(\mathbb{E}_0[C_k]\big)^2 \\
&\leq \mathbb{E}_0[C_k] + N^{-1/k}(\lambda_0\vee 1)^{2k}(2k)!\ ,
\end{aligned}$$

with

$$N^{-1/k}(\lambda_0\vee 1)^{2k}(2k)! = \exp\left[-\frac{1}{k}\log(N) + O(k\log k)\right] \to 0\ ,$$

since $k = O(\log N)^{1/4}$. Hence, $\mathrm{Var}_0[C_k] \leq \mathbb{E}_0[C_k] + o(1)$. By Chebyshev's inequality, it follows that, under $\mathbb{P}_0$,

$$C_k - E_0[C_k] \leq o(1) + O\big(\lambda_0^k/k\big)^{1/2}\ . \tag{84}$$

Let us turn to the alternative hypothesis. Let C_k^S refer to the number of k-cycles whose nodes are in S. We use the decomposition $C_k = C_k^S + (C_k - C_k^S)$. Note that $\mathbb{E}_0[C_k^S] \leq (\lambda_0 n/N)^k = o(1)$, so that $C_k = o(1)$ under $\mathbb{P}_0$ by Markov inequality. Hence, since $C_k - C_k^S$ is stochastically larger under $\mathbb{P}_S$, it follows from (84) and that, under $\mathbb{P}_S$,

$$C_k - C_k^S \geq E_0[C_k] - o(1) - O\big(\lambda_0^k/k\big)^{1/2}\ .$$

On the other hand, arguing as before, we derive that $\mathbb{E}_S[C_k^S] \sim \frac{\lambda_1^k}{2k}$ and, under $\mathbb{P}_S$,

$$C_k^S \geq \mathbb{E}_S[C_k^S] - o(1) - O\big(\lambda_1^k/k\big)^{1/2}\ .$$

Together, we find that, under $\mathbb{P}_S$,

$$C_k - E_0[C_k] \geq (1+o(1))\frac{\lambda_1^k}{2k} - o(1) - O\big(\lambda_0^k/k\big)^{1/2} - O\big(\lambda_1^k/k\big)^{1/2}\ . \tag{85}$$

since $\lambda_1 > 1$. Comparing the control under the null (84) with the control under the alternative (85), and using the fact that $\lambda_1 > \sqrt{\lambda_0}$, allows us to conclude that the test rejecting for $C_k \geq \mathbb{E}_0[C_k] + \frac{\lambda_1^k}{4k}$ is asymptotically powerful. □

Proof of Proposition 7. For $\delta = 1 + \exp[-\log^2(\log(N))]$ and $k = O(\log\log N)$,

$$g(k) = \exp\left[O(\log^3\log(N))\right] = N^{o(1)}\ ,$$

and the test runs therefore in polynomial time. In order to analyze the power of the test, we use Chebyshev's inequality and follow the tracks of Proposition 6. Under the null hypothesis,

$$\tilde{C}_k - C_k \leq (\delta-1)C_k \quad\leq\quad (\delta-1)\left[\mathbb{E}_0[C_k] + O_{P_0}\left[\sqrt{\mathbb{E}_0[C_k]}\right] + o_{\mathbb{P}_0}(1)\right] = o_{\mathbb{P}_0}(1)\ ,$$

since $(\delta-1)\,\mathbb{E}_0[C_k] \sim e^{-\log^2\log(N)}\frac{\lambda_0^k}{k} = o(1)$, as $k = O(\log\log(N))$. Under the alternative hypothesis, we derive similarly that $\tilde{C}_k \geq \delta^{-1}C_k \geq C_k - o_{\mathbb{P}_S}(1)$. Thus, we can argue as in proof of Proposition 6. □

5.4 Spectral methods

'Spectral methods' are procedures that rely heavily on an eigen-decomposition of the adjacency matrix, or related matrices like the (normalized) graph Laplacian. Since the adjacency matrix can be recovered from its spectral decomposition, any method is in principle spectral, but the term is reserved for methods that are computationally tractable (Alon, 1998; Arsić et al., 2012; Chung, 1997; Pothen et al., 1990). We explore such methods here. Let $\beta_1(\mathcal{H}) \geq \cdots \geq \beta_N(\mathcal{H})$ denote the eigenvalues of a graph $\mathcal{H}$ on N nodes, meaning the eigenvalues of its adjacency matrix. Let β_k be short for $\beta_k(\mathcal{G})$.

We first prove a simple result for the test that rejects for large values of β_1.

Proposition 8. *The test that rejects for large values of β_1 is asymptotically powerful when*

$$\lambda_1 \gg \lambda_0 \gg \sqrt{\frac{\log N}{\log \log N}} \ .$$

Proof. Krivelevich and Sudakov (2003) showed that, for $\mathcal{G} \sim \mathbb{G}(m, \lambda/m)$, $\beta_1(\mathcal{G}) \sim \sqrt{\Delta(\mathcal{G})} \vee \lambda$ in probability when the RHS goes to infinity, and when $\lambda^2 \gg \log(m)/\log\log(m)$, $\lambda \gg \sqrt{\Delta(\mathcal{G})}$ by Proposition 5 and the fact that $h^{-1}(x) \sim x/\log x$ when $x \to \infty$.

Hence, under $\mathbb{P}_0$, $\beta_1 \sim \lambda_0$, while under $\mathbb{P}_S$, $\beta_1 \geq \beta_1(\mathcal{G}_S) \sim \lambda_1$. Based on that, we conclude. □

Based on the result of Krivelevich and Sudakov (2003), and also the fact that $\beta_1 \geq \frac{2W}{N}$, we would intuit that the test based on β_1 would behave as the total degree test. However, the deviations of β_1 are rather intricate (Janson, 2005). A similar fine analysis would have to be obtained under the alternative to really understand the power properties of this test.

The test based on β_2 is particularly attractive within spectral methods because of its role in clustering. We obtain the following asymptotic performance.

Proposition 9. *The test that rejects for large values of β_2 is asymptotically powerful when $\lambda_0 \succ \log N$ and $\lambda_1 \gg \sqrt{\lambda_0} \vee (\lambda_0 \frac{n}{N})$.*

Note that $\lambda_1 \gg \lambda_0 \frac{n}{N}$ is equivalent to $p_1 \gg p_0$, which is for example true when $\limsup \alpha < 1$.

Proof of Proposition 9. We start by controlling β_2 under the null. For this, we use the following result.

Lemma 24. *When $\lambda \succ \log m$, $\beta_2(\mathbb{G}(m, \lambda/m)) = O(\sqrt{\lambda})$.*

Proof. When $\lambda \succ (\log m)^6$, the proof of Füredi and Komlós (1981) works in exactly the same way, to show that $\beta_2(\mathbb{G}(m, \lambda/m)) = O(\sqrt{\lambda})$. This was observed by Feige and Ofek (2005), who then extended the result to $\lambda \succ \log m$. □

Applying Lemma 24, we have the upper bound $\beta_2 = O(\sqrt{\lambda_0})$ under $\mathbb{P}_0$. To lower-bound β_2 under $\mathbb{P}_S$, we simply use the Courant-Fischer minimax theorem, which implies that

$$\beta_2 \geq \mathrm{I} := \min_{x \in \mathbb{R}} \frac{(\mathbf{1}_S + x\mathbf{1}_{S^c})^\top \mathbf{W} (\mathbf{1}_S + x\mathbf{1}_{S^c})}{\|\mathbf{1}_S + x\mathbf{1}_{S^c}\|^2} \ ,$$

where for $T \subset \mathcal{V} = \{1, \dots, N\}$, $\mathbf{1}_T$ denotes the vector with 1's in coordinates indexed by T, and 0's elsewhere. Straightforward calculations show that

$$\frac{1}{2}\mathrm{I} = \frac{W_{S^c}}{N-n} + \frac{W_{S,S^c}}{N-n} \min_{x \in \mathbb{R}} \frac{x + A}{x^2 + \rho} \ ,$$

where

$$W_{S,S^c} = \sum_{i \in S, j \in S^c} W_{ij} \ , \quad A := \frac{W_S}{W_{S,S^c}} - \frac{n W_{S^c}}{(N-n) W_{S,S^c}} \ , \quad \rho := \frac{n}{N-n} \ .$$

Elementary calculations show that

$$\min_{x \in \mathbb{R}} \frac{x + A}{x^2 + \rho} = -\frac{1}{2} \frac{1}{\sqrt{A^2 + \rho} + A} \ ,$$

which is increasing in A. Noting that $W_S \sim \text{Bin}(n^{(2)}, p_1)$, while $W_{S^c} \sim \text{Bin}(N^{(2)}, p_0)$ and $W_{S,S^c} \sim \text{Bin}(n(N-n), p_0)$, we see that I is stochastically increasing with λ_1, so that we may assume that $\lambda_1/\lambda_0 \to 0$. By Chebyshev's inequality, under $\mathbb{P}_S$,

$$\begin{aligned}
W_S &= \frac{n(n-1)}{2} p_1 + O(n\sqrt{p_1}) \ , \\
W_{S^c} &= \frac{(N-n)(N-n-1)}{2} p_0 + O(N\sqrt{p_0}) \ , \\
W_{S,S^c} &= n(N-n)p_0 + O(\sqrt{nNp_0}) \ ,
\end{aligned}$$

using the fact that $n = o(N)$. With this, we get

$$\begin{aligned}
A &= \frac{\frac{n^2 p_1}{2} + O(n\sqrt{p_1})}{n(N-n)p_0 + O(\sqrt{nNp_0})} - \frac{n}{N-n} \frac{\frac{(N-n)^2 p_0}{2} + O(N\sqrt{p_0})}{n(N-n)p_0 + O(\sqrt{nNp_0})} \\
&= \frac{np_1}{2(N-n)p_0} \left(1 + O(\frac{1}{n\sqrt{p_1}}) + O(\frac{1}{\sqrt{nNp_0}})\right) - \frac{1}{2} \left(1 + O(\frac{1}{N\sqrt{p_0}}) + O(\frac{1}{\sqrt{nNp_0}})\right) \\
&= \frac{\lambda_1}{2\lambda_0} + O(\frac{\lambda_1}{\lambda_0}\rho) + O(\frac{\lambda_1}{\lambda_0}\frac{1}{n\sqrt{p_1}}) - \frac{1}{2} + O(\frac{1}{\sqrt{nNp_0}}) \\
&= -\frac{1}{2} + \frac{\lambda_1}{2\lambda_0} + O(\frac{\lambda_1}{\lambda_0}\rho) + O(\frac{1}{\sqrt{n\lambda_0}}) \ ,
\end{aligned}$$

using the fact that $n = o(N)$ and then $\lambda_1 = o(\lambda_0)$.

$$-\frac{1}{2} \frac{1}{\sqrt{A^2 + \rho} + A} = -\frac{1}{2\rho}\big(\sqrt{A^2 + \rho} - A\big) = \frac{1}{2\rho}\big(\frac{\lambda_1}{\lambda_0} - 1\big) + O\big(1 + \frac{1}{\rho n \sqrt{\lambda_0}}\big) \ .$$

With this and the bounds provided by Chebyshev inequality, we find that

$$\begin{aligned}
\frac{1}{2}\mathrm{I} &= \frac{N-n}{2} p_0 + O(\sqrt{p_0}) + \big[np_0 + O(\sqrt{np_0/N})\big] \left[\frac{1}{2\rho}\big(\frac{\lambda_1}{\lambda_0} - 1\big) + O\big(1 + \frac{1}{\rho n\sqrt{\lambda_0}}\big)\right] \\
&= \frac{N-n}{2} p_0 \frac{\lambda_1}{\lambda_0} + O(np_0 + \sqrt{Np_0/n}) \\
&= \frac{\lambda_1}{2} + O(n\lambda_0/N + \sqrt{\lambda_0/n}) \ ,
\end{aligned}$$

so that $\beta_2 \geq \lambda_1 + O(n\lambda_0/N + \sqrt{\lambda_0/n})$ under $\mathbb{P}_S$. Comparing with the upper bound that we obtained for β_2 under $\mathbb{P}_0$, we conclude that, under the assumptions of Proposition 9, the test that rejects when $\beta_2 \geq \frac{1}{2}\lambda_1$ is asymptotically powerful. □

We did not explore in detail methods based on the Laplacian, or normalized Laplacian, whose eigenvalue perturbation analysis is, for example, carried in (Chung and Radcliffe, 2011).

6 Discussion

6.1 Adapting to unknown p_0 and n

In (Arias-Castro and Verzelen, 2012), we discussed in detail the case where p_0 is unknown. In this situation, the total degree test is not applicable, and we replaced it with a test based on the difference between two estimates for the degree variance. On the other hand, the scan test (based on (3)) can be calibrated in various ways without asymptotic loss of power — for example, by plugging in the estimate $\hat{p}_0 = \frac{W}{N^{(2)}}$ in place of p_0. We showed that a combination of degree variance test and the scan test are optimal when p_0 is unknown, so that the degree variance test can truly play the role of the total degree test in this situation. We believe this is the case here also. In addition to that, the broad scan test (based on (6)) can also be calibrated without asymptotic loss of power, and the same is true for all the other tests that we studied here, except for the largest connected component test in the supercritical regime.

We also discussed in (Arias-Castro and Verzelen, 2012) the case where the size of the subgraph n is unknown. This only truly affects the broad scan test, whose definition itself depends on n. As we argued in our previous paper, it suffices to apply the procedure to all possible n's, meaning, consider the multiple test based on a combination of the statistics

$$W_n^{\ddagger}, \ n = 1, \dots, N/2$$

with a Bonferroni correction. The concentration inequalities that we obtained for $W_n^{\ddagger}$ can accommodate an additional logarithmic factor that comes out of applying the union to control this statistic under $\mathbb{P}_0$, and from this we can immediately see that the test is asymptotically as powerful (up to first order).

6.2 Open problems

The cases where $\lambda_0 \to 0$ and where $\liminf \lambda_0 \geq e$ are essentially resolved. Indeed, in the first situation, the largest connected component test is asymptotically optimal by Theorem 2 and Theorem 4 case (51), while in the second situation the broad scan test is asymptotically optimal by Theorem 1 and Theorem 4 cases (53) and (54), together with Theorem 5. The case where $0 < \lambda_0 < e$ is fixed is not completely resolved. Since the triangle test has non-negligible power as soon as λ_1 is bounded away from 0, consider τ defined as the largest real such that no test for $\mathbb{G}(N, \frac{\lambda_0}{N})$ versus $\mathbb{G}(N, \frac{\lambda_0}{N}; n, \frac{\lambda_1}{n})$ is asymptotically powerful when $\limsup \lambda_1 < \tau$. Theorems 2 and 3 provide some upper bounds on τ.

Open problem 1. *Compute τ as a function of λ_0 and $\kappa := \limsup \frac{\log n}{\log N}$.*

Although we proved that the broad scan test was asymptotically optimal when $\liminf \lambda_0 \geq e$, its performance was described only indirectly in terms of λ_1 in the case (13).

Open problem 2. *Compute, as a function of λ_1 and possibly κ (defined above), the limits inferior and superior of*

$$\sup_{k=n/u_N}^{n} \frac{\mathbb{E}_S[W^*_{k,S}]}{k} \, .$$

We also formulate an open problem that connects directly with the planted clique problem. We saw that the broad scan test is powerful when λ_1 is sufficiently large, but we do not know how to compute it in polynomial time. Is there a polynomial-time test that can come close to that?

Open problem 3. *Find a polynomial-time test that is asymptotically powerful for testing $\mathbb{G}(N, p_0)$ versus $\mathbb{G}(N, p_0; n, p_1)$ when $n^2/N = O(1)$, while $\lambda_0 \to \infty$ and $\lambda_1 = O(1)$.*

7 Proofs of auxiliary results

7.1 Proof of Lemma 6

Fix $\epsilon > 0$ and define $x := 2\left[(1+\epsilon) + \sqrt{(1+\epsilon)^2 + \lambda_1(1+\epsilon)}\right]$. First, we control the deviations of $W^*_{k,S}$. Define $q_k = (\lambda_1 + x)/(k-1)$ and notice that $q_k \geq p_1$ for $n/u_N \leq k \leq n$. Since $\log(1+t) \leq t$ for any $t > -1$, we have

$$H_{p_1}(q_k) := q_k \log\left(\frac{q_k}{p_1}\right) + (1-q_k)\log\left(\frac{1-q_k}{1-p_1}\right) \geq q_k \log\left(\frac{q_k}{p_1}\right) - q_k + p_1 \ .$$

Applying an union bound and Chernoff inequality (10), we control the deviations of $W^*_{k,S}$:

$$\mathbb{P}_S\left[W^*_{k,S} \geq k^{(2)} q_k\right] \leq \binom{n}{k} \exp\left[-k^{(2)} H_{p_1}(q_k)\right] \leq \exp[kA_k] \ ,$$

where

$$A_k := \log\left(\frac{en}{k}\right) - \frac{k-1}{2}\left(q_k \log\left(\frac{q_k}{p_1}\right) - q_k + p_1\right) \ .$$

Observe that x is larger than 2. As a consequence, we obtain

$$\begin{aligned} A_k &= 1 + \log\left(\frac{n}{k}\right) - \frac{\lambda_1 + x}{2}\log\left(\frac{n(\lambda_1+x)}{(k-1)\lambda_1}\right) + \frac{\lambda_1+x}{2} - \frac{\lambda_1(k-1)}{2n} \\ &\leq 1 + \frac{x}{2} - \frac{\lambda_1+x}{2}\log\left(\frac{\lambda_1+x}{\lambda_1}\right) - \frac{\lambda_1}{2}\left[\frac{k-1}{n} - 1 - \log\left(\frac{k-1}{n}\right)\right] \\ &\leq 1 - \frac{x^2}{4(\lambda_1+x)} \ , \end{aligned}$$

where we used in the last line the inequalities $t - \log t - 1 \geq 0$ and $\log(1-t) \leq -t - t^2/2$, valid for any $t \geq 0$. By definition of x, we have $x^2/(4(\lambda_1+x)) = 1+\epsilon$. In conclusion, we have proved that for any integer k between n/u_N and n

$$\mathbb{P}_S\left[\frac{W^*_{k,S}}{k} \geq \frac{\lambda_1+x}{2}\right] \leq \exp\left[-k\epsilon\right] \ . \tag{86}$$

Let us now control the lower deviations of $\frac{1}{k}W^*_{k,S}$ using Lemma 7

$$\mathbb{P}_S\left[\frac{W^*_{k,S}}{k} \leq \mathbb{E}_S\left[\frac{W^*_{k,S}}{k}\right] - \left(\mathbb{E}_S\left[\frac{W^*_{k,S}}{k}\right]\right)^{1/2}\frac{8}{k^{1/2}}\right] \leq 2^{-8} \ .$$

For k large enough, $\exp\left[-k\varepsilon\right] \leq 1/2$, which therefore implies that

$$\mathbb{E}_S\left[\frac{W^*_{k,S}}{k}\right] \leq \left(\mathbb{E}_S\left[\frac{W^*_{k,S}}{k}\right]\right)^{1/2}\frac{8}{(n/u_N)^{1/2}} + \frac{\lambda_1+x}{2} \ ,$$

since $k \geq n/u_N$. Taking the supremum over k and letting n go to infinity, we conclude that

$$\liminf \bigvee_{k=n/u_N}^{n} \mathbb{E}_S\left[\frac{W^*_{k,S}}{k}\right] \leq \liminf \frac{\lambda_1+x}{2} = \liminf \frac{\lambda_1}{2} + (1+\epsilon) + \sqrt{(1+\epsilon)^2 + \lambda_1(1+\epsilon)} \ .$$

Then letting ϵ going to zero allows us to conclude.

7.2 Some combinatorial results

We state and prove some combinatorial results.

Lemma 25 (Extension of Cayley's identity)**.** *The number $T_k^{(\ell)}$ of labelled trees of size ℓ containing a given labelled tree of size k satisfies*

$$T_k^{(\ell)} = k\ell^{\ell-k-1} \ .$$

The number $T_{k_1,\dots k_r}^{(\ell)}$ of labelled trees of size ℓ containing a given labelled forest with tree components of size $k_1,\dots,k_r$ satisfies

$$T_{k_1,\dots,k_r}^{(\ell)} \leq \left(\frac{k}{r}\right)^r \ell^{\ell-k+r-1}(\ell-k+r-1)^{r-1} \ ,$$

with $k = \sum_{i=1}^r k_i$.

Proof. The proof relies on the double counting argument of Pitman (Aigner and Ziegler, 2010). Noting $\mathcal{T}$ the fixed tree of size k, we count in two ways the number of labelled trees of size ℓ that contain $\mathcal{T}$ and whose vertices outside $\mathcal{T}$ have been ordered. Straightforwardly, we have $T_k^{(\ell)}(\ell-k)!$ such trees. Alternatively, we consider the following way of building such a labelled ordered tree:

1. Start from $\mathcal{T}$.

2. Choose any vertex $\tilde{u}_0$ among the original tree $\mathcal{T}$ and any vertex $\tilde{v}_0$ among the $(\ell-k)$ remaining vertices. Add an edge between $\tilde{u}_0$ and $\tilde{v}_0$. Root the given tree — now of size $k+1$ — at $\tilde{v}_0$. Consider all the $\ell-k-1$ remaining vertices as rooted trees of size 1.

3. Then, perform the iterative construction of Pitman. At each step $i = 1,\dots,\ell-k-1$, add an edge in the following way: choose any starting vertex u_i among the ℓ vertices and note ρ_i the root of the tree containing u_i. Choose any ending vertex v_i among the $(\ell-k-i)$ roots other than ρ_i. This so-obtained tree is rooted at ρ_i.

4. Let $v_{\ell-k}$ denote the root of the final tree.

All in all, we have $k\ell^{\ell-k-1}(\ell-k)!$ such constructions and the sequence $v_1,\dots,v_{\ell-k}$ obtained in Step 3 provides an ordering for the vertices not in $\mathcal{T}$.

Lemma 26. *For any labelled tree $\tilde{\mathcal{T}}$ of size ℓ that contains $\mathcal{T}$ and whose vertices outside $\mathcal{T}$ have been ordered, there exists one and only one construction of $\tilde{\mathcal{T}}$ based on the algorithm above.*

Comparing the two counts leads to the desired result.

Proof of Lemma 26. Let us slightly modify the iterative construction of Pitman by putting an orientation on the added edges: the first edge is oriented from $\tilde{v}_0$ to $\tilde{u}_0$. For any $i = 1,\dots,\ell-k-1$, the edge between u_i and v_i is oriented from u_i to v_i. The so-obtained partially oriented tree is noted $\overrightarrow{\mathcal{T}}_{u,v}$.

Observe that except for $v_{\ell-k}$ which has no parents, all other nodes v_i have one and only one parent. Also, observe that except for the edge $\tilde{v}_0 \to \tilde{u}_0$, all edges between nodes in the subtree $\mathcal{T}$ and nodes in $\{v_1,\dots v_{\ell-k}\}$ leave the subtree $\mathcal{T}$. By a simple induction, this leads us to the following claim:

Claim 1: All partially oriented tree $\overrightarrow{\mathcal{T}}_{u,v}$ based on Pitman construction with sequences $(u,v) = (\tilde{u}_0, \tilde{v}_0, u_1, \ldots, u_{\ell-k-1}, v_1, \ldots, v_{\ell-k-1}, v_{\ell-k})$ satisfy the following property

$$(P) \quad \begin{cases} \text{Any edge in } \mathcal{T} \text{ is undirected,} \\ \text{Any edges on the unique path between } v_{\ell-k} \text{ and } \mathcal{T} \text{ is oriented towards } \mathcal{T}, \\ \text{Any other edge (not in } \mathcal{T}\text{) is oriented in the opposite direction to } \mathcal{T}. \end{cases}$$

In fact, this property characterizes the oriented partially trees $\overrightarrow{\mathcal{T}}_{u,v}$.

Claim 2: Conversely, for any sequence $v = (v_1, \ldots, v_{\ell-k})$ and any partially oriented tree $\overrightarrow{\mathcal{T}}$ of size ℓ satisfying (P), there exists a unique sequence, $(\tilde{u}_0, \tilde{v}_0, u_1, \ldots, u_{\ell-k-1})$ such that $\overrightarrow{\mathcal{T}}_{u,v} = \mathcal{T}$.

Proof of Claim 2. The uniqueness is straightforward. Given $\overrightarrow{\mathcal{T}}$, define $\tilde{u}_0$ as the unique child in $\mathcal{T}$ and define $\tilde{v}_0$ the parent of $\tilde{u}_0$. For any $i = 1, \ldots, \ell-k-1$, denote u_i the parent of v_i. These sequences are lawful for the Pitman construction. Indeed, at step i, v_i is not in the same connected component as u_i and v_i is still a root of a connected component.

Then, the lemma proceeds from the fact that for any tree $\mathcal{T}$ of size ℓ and any sequence $v = (v_1, \ldots, v_{\ell-k})$, there exists one and only partially orientation of $\mathcal{T}$ satisfying (P). $\square$

We now prove the second part of Lemma 25, relying on the same double counting argument. Write $k = k_1 + \ldots + k_r$. Noting $\mathcal{F}$ the fixed forest with (labelled) connected components $\mathcal{T}_1, \ldots, \mathcal{T}_r$ of respective sizes $k_1, \ldots, k_r$, we count in two ways the number of labelled trees of size ℓ that contain $\mathcal{F}$ and whose vertices outside $\mathcal{F}$ have been ordered. Straightforwardly, we have $T^{(\ell)}_{k_1,\ldots,k_r}(\ell-k)!$ such trees. Alternatively, we consider the following Pitman construction:

1. Start from $\mathcal{F}$.
2. For any $j = 1, \ldots, r$, choose any vertex $w_j \in \mathcal{T}_j$. Root $\mathcal{T}_j$ at w_j. Consider all the $\ell - k$ remaining vertices as rooted trees of size 1.
3. Then, perform the iterative construction of Pitman: at each step $i = 1, \ldots, \ell - k + r - 1$, add an edge in the following way: choose any starting vertex u_i among the ℓ vertices and note ρ_i the root of the tree containing u_i. Choose any ending vertex v_i among the remaining $(\ell - k + r - i)$ roots other than ρ_i. The resulting tree is rooted at ρ_i.
4. Let $v_{\ell-k+r}$ denote the root of the final tree.

All in all, we have $\big(\prod_{j=1}^r k_j\big)\ell^{\ell-k+r-1}(\ell-k+r-1)!$ such constructions. And the sequence $v_1, \ldots, v_{\ell-k+r}$ obtained in Step 3 provides an ordering of the vertices outside $\mathcal{F}$ if we ignore the w_j's in that sequence.

Lemma 27. *Any tree that contains $\mathcal{F}$ and whose vertices outside $\mathcal{F}$ are ordered by the sequence $(t_1, \ldots, t_{\ell-k})$ can be constructed in this way.*

Consequently we have

$$T^{(\ell)}_{k_1,\ldots,k_r}(\ell-k)! \leq \big(\prod_{i=1}^r k_i\big)\, \ell^{\ell-k+r-1}(\ell-k+r-1)! \ ,$$

from which we derive the (crude) bound

$$T^{(\ell)}_{k_1,\ldots,k_r} \leq \left(\frac{k}{r}\right)^r \ell^{\ell-k+r-1}(\ell-k+r-1)^{r-1} \ .$$

$\square$

Proof of Lemma 27. Consider a tree $\mathcal{T}^{\ell}$ that contains $\mathcal{F}$ and whose vertices outside $\mathcal{F}$ are ordered in the following sequence $(t_1, \ldots, t_{\ell-k})$.

Claim. There exists a (non-necessarily unique) orientation of the edges outside $\mathcal{F}$ such that any node in $t_1, \ldots, t_{\ell-k-1}$ has exactly one parent, $t_{\ell-k}$ has no parent, and any tree $\mathcal{T}_i$ in $\mathcal{F}$ has exactly one parent.

Proof of Claim 1: Collapse each of the trees $\mathcal{T}_i$ into a single node, to obtain the tree $\mathcal{T}^{\ell-k+r}$ with $\ell - k + r$ nodes. Then, we prove the result for $\mathcal{T}^{\ell-k+r}$ by a simple induction on the number of nodes.

For any $i \in \{1, \ldots, r\}$, define ω_i as the unique node in $\mathcal{T}_i$. We define the sequence $v := (\omega_1, \ldots, \omega_r, t_1, \ldots t_{\ell-k})$. Finally, we define u_i as the unique parent of v_i for any $i \leq \ell - k + r - 1$. It is straightforward to check that these sequences ω, v and u are lawful for the Pitman construction and allow to build $\mathcal{F}$. $\square$

7.3 Proof of Lemma 14

By definition,

$$\mathrm{Var}_0[N_k^{\mathrm{tree}}] = \sum_{C_1, C_2} \left(\mathbb{P}_0[\mathcal{G}_{C_1} \text{ and } \mathcal{G}_{C_2} \text{ are trees}] - \mathbb{P}_0[\mathcal{G}_{C_1} \text{ is a tree}]\mathbb{P}_0[\mathcal{G}_{C_2} \text{ is a tree}]\right) ,$$

where the sum ranges over subsets C_1, C_2 of size k.

In the sequel, we let $q = |C_1 \cap C_2|$ and let r denote the number of connected components of $C_1 \cap C_2$. Note that, when $q = 0$, the corresponding terms in the sum above are zero. When $q \geq 1$, we define

$$B_{r,q} = \mathbb{P}_0\left[\mathcal{G}_{C_1}, \mathcal{G}_{C_2} \text{ are trees and } \mathcal{G}_{C_1 \cap C_2} \text{ has } r \text{ connected components}\right] ,$$

so that

$$\mathbb{P}_0[\mathcal{G}_{C_1} \text{ and } \mathcal{G}_{C_2} \text{ are trees}] = \sum_{r=1}^{q} B_{r,q} .$$

Note that $\mathcal{G}_{C_1 \cap C_2}$ is a forest when $\mathcal{G}_{C_1}$ and $\mathcal{G}_{C_2}$ are trees.

We derive $B_{1,q}$ first. Under the event $\{\mathcal{G}_{C_1}, \mathcal{G}_{C_2} \text{ and } \mathcal{G}_{C_1 \cap C_2} \text{ are trees}\}$, there are exactly $2k-1-q$ edges in $\mathcal{G}_{C_1} \cup \mathcal{G}_{C_2}$ among the potential $2k^{(2)} - q^{(2)}$ edges. Let us count the number of configurations compatible with this event. By Cayley's identity, there are q^{q-2} configurations for the tree $\mathcal{G}_{C_1 \cap C_2}$. The tree $\mathcal{G}_{C_1 \cap C_2}$ being fixed, we apply Lemma 25 to derive that there are qk^{k-q-1} configurations for $\mathcal{G}_{C_1}$ and qk^{k-q-1} configurations for $\mathcal{G}_{C_2}$. All in all, we get

$$B_{1,q} \quad = \quad q^{q-2}[qk^{k-q-1}]^2 p_0^{2k-1-q}(1-p_0)^{2k^{(2)}-q^{(2)}-2k+1+q} .$$

Then, we upper bound $B_{r,q}$ for $r \geq 2$. Under the event defined in $B_{r,q}$, there are $2k - 2 - q + r$ edges in $\mathcal{G}_{C_1} \cup \mathcal{G}_{C_2}$ among the potential $2k^{(2)} - q^{(2)}$ edges. By Lemma 19 , there are less than q^{q-2} configurations for the forest $\mathcal{G}_{C_1 \cap C_2}$ with r connected components. $\mathcal{G}_{C_1 \cap C_2}$ being fixed, Lemma 25 tells us that there are less than $\left(\frac{q}{r}\right)^r k^{k-q+r-1}(k-q+r-1)^{r-1}$ possible configurations to complete $\mathcal{G}_{C_1}$ and (independently) for $\mathcal{G}_{C_2}$. It then follows that

$$\begin{aligned} B_{r,q} &\leq q^{q-2}\left[\left(\frac{q}{r}\right)^r k^{k-q+r-1}(k-q+r-1)^{r-1}\right]^2 p_0^{2k-q+r-2}(1-p_0)^{2k^{(2)}-q^{(2)}-2k+q-r+2} \\ &\leq B_{1,q}\left(\frac{p_0}{1-p_0}\right)^{r-1} k^{6r-4} , \end{aligned}$$

using the fact that $q \le k$. Summing over r leads to

$$\sum_{r=2}^{q} B_{r,q} \le B_{1,q}\, k^2 \sum_{r=2}^{q} \left[\frac{p_0 k^6}{1-p_0}\right]^{r-1} \le B_{1,q}\, \frac{p_0 k^8}{1-p_0-p_0k^6} = o\big(B_{1,q}\big) \ ,$$

since $p_0 k^8 = o(1)$. Thus, when $|C_1 \cap C_2| = q \ge 1$, we obtain

$$\begin{aligned} \mathbb{P}_0[\mathcal{G}_{C_1} \text{ and } \mathcal{G}_{C_2} \text{ are trees}] &= B_{1,q} + o\big(B_{1,q}\big) \\ &\prec q^q k^{2k-2q-2} p_0^{2k-1-q} \ . \end{aligned}$$

We can now bound the variance. The number of subsets (C_1, C_2) of size k such that $C_1 \cap C_2 = q$ equals $\binom{N}{k}\binom{k}{q}\binom{N-k}{k-q}$. Thus, we derive

$$\begin{aligned} \mathrm{Var}_0[N_k^{\text{tree}}] &\prec \sum_{q=1}^{k} \binom{N}{k}\binom{k}{q}\binom{N-k}{k-q} q^q k^{2k-2q-2} p_0^{2k-q-1} \\ &\prec N \sum_{q=1}^{k} \frac{q^q k^{2k-2q-2}}{q!(k-q)!^2} \lambda_0^{2k-2q-1} \\ &\prec N \sum_{q=1}^{k} \frac{(\lambda_0 e)^k}{\lambda_0 k^2} A_{k-q}, \qquad A_\ell := \left(\frac{k\sqrt{\lambda_0 e}}{\ell}\right)^{2\ell} \ , \end{aligned}$$

by Stirling's lower bound. By convention, $A_0 = 1$. The function $\ell \to A_\ell$ is easily seen to be increasing over $(0, k\sqrt{\lambda_0/e})$ and decreasing over $(k\sqrt{\lambda_0/e}, \infty)$. Thus, when $\lambda_0 < e$, we have $A_{k-q} \le A_{k\sqrt{\lambda_0/e}}$; and when $\lambda_0 > e$, we have $A_{k-q} \le A_k$; this is for all $q = 1, \ldots, k$. Then summing over q, we obtain the stated bounds in each case.

7.4 Proof of Lemma 15

First, we deal with the expectation.

$$\mathbb{E}_S[N_{k,S,q}^{\text{tree}}] = \sum_{C_1 \subset S, |C_1| = q} \ \sum_{C_2 \subset S^c, |C_2| = k-q} \mathbb{P}_S[\mathcal{G}_{C_1} \text{ and } \mathcal{G}_{C_1 \cup C_2} \text{ are trees}] \ .$$

When $\mathcal{G}_{C_1}$ and $\mathcal{G}_{C_1 \cup C_2}$ are both trees, there are $q-1$ edges in $\mathcal{G}_{C_1}$ and $k-q$ additional edges in $\mathcal{G}_{C_1 \cup C_2}$. The number of configurations for $\mathcal{G}_{C_1}$ is q^{q-2} (Cayley's Identity). By Lemma 25, when $\mathcal{G}_{C_1}$ is fixed, there remains qk^{k-q-1} possible configurations for $\mathcal{G}_{C_1 \cup C_2}$. As for the previous variance computation, we apply to control this probability. Hence, we get

$$\begin{aligned} \mathbb{P}_S[\mathcal{G}_{C_1} \text{ and } \mathcal{G}_{C_1 \cup C_2} \text{ are trees}] &= q^{q-2} q k^{k-q-1} p_1^{q-1} p_0^{k-q} (1-p_1)^{q^{(2)}-q+1} (1-p_0)^{k^{(2)}-q^{(2)}-k+q} \\ &\succ q^{q-1} k^{k-q-1} p_1^{q-1} p_0^{k-q} \ , \end{aligned}$$

since $(q^{(2)} - q + 1)p_1 \le k^2 p_1 \asymp k^2/n = o(1)$ and $(k^{(2)} - q^{(2)} - k + q)p_0 \le k^2 p_0 \asymp k^2/N = o(1)$. Hence, using the fact that $nk = o(N)$ and the usual bound $m! \le \sqrt{m}(m/e)^m$, and we derive

$$\begin{aligned} \mathbb{E}_S[N_{k,S,q}^{\text{tree}}] &\succ \binom{n}{q}\binom{N-n}{k-q} q^{q-1} k^{k-q-1} p_1^{q-1} p_0^{k-q} \\ &\succ n \frac{q^{q-1} k^{k-q-1} \lambda_1^{q-1} \lambda_0^{k-q}}{q!(k-q)!} \\ &\succ n \frac{(e\lambda_1)^k}{\lambda_1 k^3} \left(\frac{\lambda_0 k}{\lambda_1 (k-q)}\right)^{k-q} \ . \end{aligned}$$

This quantity is maximized with respect to q when $(k-q)/k = \lambda_0/(\lambda_1 e)$, and taking $q := k - \lfloor \frac{\lambda_0}{\lambda_1 e} k \rfloor$ leads to

$$\mathbb{E}_S[\tilde{N}^T_{k,S,q}] \succ n\frac{(e\lambda_1)^k}{\lambda_1 k^3}\exp\left(\frac{\lambda_0}{\lambda_1 e}k\right) \ .$$

Let us turn to the variance. Again, we decompose it as a sum over $(C_1, C_2) \subset S^2$ and $(C_3, C_4) \subset (S^{\mathsf{c}})^2$ depending on the sizes $s = |C_1 \cap C_2|$ and $r = |C_3 \cap C_4|$. By independence of the edges, only the subsets such $(r, s) \neq (0, 0)$ play a role in the variance. We have

$$\begin{aligned}
\mathrm{Var}_S[\tilde{N}^T_{k,S,q}] &\leq \sum_{s=1}^{q}\sum_{r=0}^{k-q}\sum_{|C_1\cap C_2|=s}\ \sum_{|C_3\cap C_4|=r}\mathbb{P}_S[\mathcal{G}_{C_1},\ \mathcal{G}_{C_2},\ \mathcal{G}_{C_1\cup C_3},\ \mathcal{G}_{C_2\cup C_4}\text{ are trees}]\\
&\quad+\sum_{r=2}^{k-q}\ \sum_{|C_1\cap C_2|=0}\ \sum_{|C_3\cap C_4|=r}\mathbb{P}_S[\mathcal{G}_{C_1},\ \mathcal{G}_{C_2},\ \mathcal{G}_{C_1\cup C_3},\ \mathcal{G}_{C_2\cup C_4}\text{ are trees}]\\
&= B_1 + B_2 \ .
\end{aligned}$$

First we consider the sum B_1 where r is positive. Therefore, fix $C_1, C_2 \subset S$ and $C_3, C_4 \subset S^{\mathsf{c}}$ with $|C_1| = |C_2| = q$, $|C_3| = |C_4| = k - q$, $|C_1 \cap C_2| = s \geq 1$ and $|C_3 \cap C_4| = r \geq 1$, and for $1 \leq t_1 \leq s$ and $1 \leq t_2 \leq r + s$, define

$$\begin{aligned}
\mathcal{A}_{(t_1,t_2)} &:= \{\mathcal{G}_{C_1},\ \mathcal{G}_{C_2},\ \mathcal{G}_{C_1\cup C_3},\ \mathcal{G}_{C_2\cup C_4}\text{ are trees}\}\\
&\quad\ \cap\{\mathcal{G}_{C_1\cap C_2}\text{ has } t_1 \text{ connected components}\}\\
&\quad\ \cap\left\{\mathcal{G}_{(C_1\cap C_2)\cup(C_3\cap C_4)}\text{ has } t_2 \text{ connected components}\right\} \ .
\end{aligned}$$

(The dependency of $\mathcal{A}_{(t_1,t_2)}$ on C_1, C_2, C_3, C_4 is left implicit.)

We first control $\mathbb{P}_S[\mathcal{A}_{(1,1)}]$. Under the event $\mathcal{A}_{(1,1)}$, the graph $\mathcal{G}_{C_1} \cup \mathcal{G}_{C_2}$ contains $2q - 1 - s$ edges and the graph $\mathcal{G}_{C_1\cup C_3} \cup \mathcal{G}_{C_2\cup C_4}$ contains $2(k - q) - r$ additional edges. Indeed, the number of edges in the last graph is equal to

$$(|C_1\cup C_3|-1)+(|C_2\cup C_4|-1)-(|(C_1\cap C_2)\cup(C_3\cap C_4)|-1) = k-1+k-1-(r+s-1) = 2k-r-s-1$$

Applying Lemma 25, there are s^{s-2} possible configurations for $\mathcal{G}_{C_1\cap C_2}$ and then $[sq^{q-s-1}]^2$ possible configurations to complete $\mathcal{G}_{C_1} \cup \mathcal{G}_{C_2}$. The graph $\mathcal{G}_{C_1} \cup \mathcal{G}_{C_2}$ been fixed, there are $s(s + r)^{r-1}$ configurations for $\mathcal{G}_{(C_1\cap C_2)\cup(C_3\cap C_4)}$, since this is a tree with $s + r$ nodes containing the given tree $\mathcal{G}_{C_1\cap C_2}$ with s nodes. By the same token, $\mathcal{G}_{C_1\cup C_3}$ is a tree with k nodes that includes the given tree $\mathcal{G}_{C_1\cup(C_3\cap C_4)}$ with $q + r$ nodes, and similarly for $\mathcal{G}_{C_2\cup C_4}$, there at most $[(q + r)k^{k-q-r-1}]^2$ possible configurations to complete $\mathcal{G}_{C_1\cup C_3} \cup \mathcal{G}_{C_2\cup C_4}$. Thus, we obtain

$$\mathbb{P}_S[\mathcal{A}_{(1,1)}] \leq s^{s-2}[sq^{q-s-1}]^2 s(s+r)^{r-1}[(q+r)k^{k-q-r-1}]^2 p_1^{2q-1-s} p_0^{2(k-q)-r} =: A_{1,1} \ . \tag{87}$$

Let us now control the probability of $\mathcal{A}_{(t_1,t_2)}$ for t_1 or t_2 strictly larger than one. First, observe that whenever $t_2 < t_1$, $\mathcal{A}_{t_1,t_2}$ is empty. Indeed, if $t_2 < t_1$, there is a path in $C_3 \cap C_4$ between two connected components of $\mathcal{G}_{C_1\cap C_2}$. However, these two connected components are also related by a different (since $C_1 \cap C_3 = \emptyset$) path in C_1 (since $\mathcal{G}_{C_1}$ is a tree), and that contradicts the fact that $\mathcal{G}_{C_1\cup C_3}$ is a tree. Hence, we may assume that $t_2 \geq t_1$. By Lemmas 25 and 19, there are at most s^{s-2} possible configurations for the forest $\mathcal{G}_{C_1\cap C_2}$, and when this is fixed, there are at most

$$\left[(s/t_1)^{t_1} q^{q-s+t_1-1}(q-s+t_1-1)^{t_1-1}\right]^2 \leq [sq^{q-s}q^{3(t_1-1)}]^2$$

possible configurations to complete $\mathcal{G}_{C_1} \cup \mathcal{G}_{C_2}$. With $\mathcal{G}_{C_1} \cup \mathcal{G}_{C_2}$ being fixed, the number of possible configurations for $\mathcal{G}_{(C_1 \cap C_2) \cup (C_3 \cap C_4)}$ is at most the number of trees that contain $\mathcal{G}_{C_1 \cap C_2}$ — which is at most

$$(s/t_1)^{t_1}(s+r)^{s+r-s+t_1-1}(s+r-s+t_1-1)^{t_1-1} \leq s^{t_1}(s+r)^{r+2t_1-2}$$

by Lemma 25 — times k^{t_2-1}, which bounds the number of ways of erasing t_2-1 edges in this tree to obtain a forest with t_2 components. The graph $\mathcal{G}_{C_1 \cup (C_3 \cap C_4)}$ contains t_2-t_1+1 connected components. By Lemma 25, there are no more than

$$\left[\left(\frac{q+r}{t_2-t_1+1}\right)^{t_2-t_1+1} k^{k-q-r+t_2-t_1}(k-q-r+t_2-t_1)^{t_2-t_1}\right]^2 \leq [(q+r)k^{k-r-q}k^{3(t_2-t_1)}]^2$$

possible configurations to complete $\mathcal{G}_{C_1 \cup C_3} \cup \mathcal{G}_{C_2 \cup C_4}$. The number of edges in $\mathcal{G}_{C_1} \cup \mathcal{G}_{C_2}$ is $2(q-1)-(s-t_1)$, while the number of edges in $\mathcal{G}_{C_1 \cup C_3} \cup \mathcal{G}_{C_2 \cup C_4}$ is

$$2(k-q)-(r-(t_2-t_1)) = 2(k-q)-r+t_2-t_1 \ .$$

All together, and with some elementary simplifications, we arrive at the following bound

$$\mathbb{P}_S[\mathcal{A}_{(t_1,t_2)}] \ \leq \ A_{1,1}k^{7(t_2-t_1)+9t_1-6}p_1^{t_1-1}p_0^{t_2-t_1} \ .$$

Since $k^{O(1)}(p_0+p_1) = o(1)$, it follows that

$$\sum_{t_1=1}^{s}\sum_{t_2=1}^{r+s} \mathbb{P}_S(\mathcal{A}_{t_1,t_2}) \prec A_{1,1}.$$

Using the definition of $A_{1,1}$ in (87) and the definition of q, we bound B_1

$$\begin{aligned}
B_1 &\prec \sum_{s=1}^{q}\sum_{r=0}^{k-q}\binom{n}{q}\binom{N-n}{k-q}\binom{q}{s}\binom{k-q}{r}\binom{n-q}{q-s}\binom{N-n-k+q}{k-q-r}A_{1,1} \\
&\prec n\sum_{s,r}\frac{s^{s+1}q^{2(q-s-1)}(s+r)^{r-1}k^{2(k-q-r)-2}(q+r)^2}{r!s!(q-s)!^2(k-q-r)!^2}\lambda_1^{2q-s-1}\lambda_0^{2(k-q)-r} \\
&\prec n\sum_{r,s}e^{2k-r-s}\left(\frac{s+r}{r}\right)^r\left(\frac{q}{q-s}\right)^{2(q-s)}\left(\frac{k}{k-q-r}\right)^{2(k-q-r)}\lambda_1^{2q-s-1}\lambda_0^{2(k-q)-r+1} \\
&\prec \frac{n}{\lambda_1}\sum_{r,s}e^{4k-2q-3r-s}\left(\frac{s+r}{r}\right)^r\left(\frac{q}{q-s}\right)^{2(q-s)}\left(\frac{k-q}{k-q-r}\right)^{2(k-q-r)}\lambda_1^{2k-2r-s-1}\lambda_0^{r} \\
&\prec \frac{n}{\lambda_1}\sum_{r,s}e^{4k-2q-3r}\left(\frac{q}{q-s}\right)^{2(q-s)}\left(\frac{k-q}{k-q-r}\right)^{2(k-q-r)}\lambda_1^{2k-2r-s-1}\lambda_0^{r} \ .
\end{aligned}$$

We have applied Stirling's lower bound in the fourth line; we haved used the definition of q to control $k/(k-q)$ in the fifth line

$$\left(\frac{k}{k-q}\right)^{2(k-q-r)} = \left(\frac{k}{\lfloor\frac{\lambda_0 k}{\lambda_1 e}\rfloor}\right)^{2(k-q-r)} \leq \left(\frac{\lambda_1 e}{\lambda_0}\right)^{2(k-q-r)}\left(1-\frac{\lambda_1 e}{k\lambda_0}\right)^{-2k} = O(1)\left(\frac{\lambda_1 e}{\lambda_0}\right)^{2(k-q-r)} \quad ;$$

and we have upper-bounded $(1+s/r)^r$ by e^s in the last line. Note that

$$e^{-3r}\left(\frac{k-q}{k-q-r}\right)^{2(k-q-r)}\lambda_1^{-2r}\lambda_0^r = \left(\frac{(k-q)e^{3/2}\frac{\lambda_1}{\sqrt{\lambda_0}}}{k-q-r}\right)^{2(k-q-r)}\left(e^{3/2}\frac{\lambda_1}{\sqrt{\lambda_0}}\right)^{-2(k-q)}$$

is decreasing with respect to r since $\lambda_1^2 e > \lambda_0$. As a consequence, we have

$$B_1 \quad \prec \quad \frac{nk}{\lambda_1}\sum_{\ell=1}^{q} e^{4k-2q}\lambda_1^{2k-q}D_\ell \qquad D_\ell := \left(\frac{q}{\ell}\right)^{2\ell}\lambda_1^\ell$$

The function $\ell \to D_\ell$ is easily seen to be maximized at $\ell = q\sqrt{\lambda_1}/e$. This allows us to conclude that

$$B_1 \quad \prec \quad \frac{nk^2}{\lambda_1}e^{4k-2q+2q\sqrt{\lambda_1}/e}\lambda_1^{2k-q} \; .$$

Finally, we bound B_2 following a similar strategy. First, we observe that the probability of the event $\mathcal{B} := \{\mathcal{G}_{C_1},\, \mathcal{G}_{C_2},\, \mathcal{G}_{C_1\cup C_3},\, \mathcal{G}_{C_2\cup C_4}$ are trees$\}$ is equivalent to the probability of the event $\mathcal{B}_1 := \mathcal{B}\cap\{\mathcal{G}_{C_3\cap C_4}$ is a tree$\}$. This follows from the fact that the event $\mathcal{B}_r := \mathcal{B}\cap\{\mathcal{G}_{C_3\cap C_4}$ contains r trees$\}$ involves $r-1$ more edges than $\mathcal{B}_1$ while the number of possible configurations in $\mathcal{B}_r$ is does not increase more than by a factor $k^{O(1)r}$ compared to $\mathcal{B}_1$.

$$\begin{aligned}
B_2 \quad &= \quad \sum_{r=2}^{k-q}\sum_{|C_1\cap C_2|=0}\sum_{|C_3\cap C_4|=r}\mathbb{P}_S[\mathcal{G}_{C_1},\, \mathcal{G}_{C_2},\, \mathcal{G}_{C_1\cup C_3},\, \mathcal{G}_{C_2\cup C_4} \text{ are trees}]\\
&\prec \quad \sum_{r=2}^{k-q}\sum_{|C_1\cap C_2|=0}\sum_{|C_3\cap C_4|=r}\mathbb{P}_S[\mathcal{G}_{C_1},\, \mathcal{G}_{C_2},\, \mathcal{G}_{C_1\cup C_3},\, \mathcal{G}_{C_2\cup C_4}\text{ , and } \mathcal{G}_{C_3\cap C_4} \text{ are trees}]\\
&\prec \quad \sum_{r=2}^{k-q} r^{r-2}\left(q^{q-2}\right)^2\left(\left(\frac{q+r}{2}\right)^2 k^{k-q-r+1}(k-q-r+1)\right)^2 p_1^{2(q-1)}p_0^{2(k-q)-r+1}\\
&\qquad \times\binom{n}{q}^2\binom{N-n}{r}\binom{N-n}{k-q-r}^2\\
&\prec \quad \sum_{r=2}^{k-q}\frac{n^2}{N}r^2k^4\lambda_1^{2(q-1)}\lambda_0^{2(k-q)-r+1}e^{2k-r}\left(\frac{k}{k-q-r}\right)^{2(k-q-r)}\\
&\prec \quad k^6\sum_{r=2}^{k-q}\frac{n^2}{N}\lambda_1^{2k-2r-2}\lambda_0^{r+1}e^{4k-2q-3r}\left(\frac{k-q}{k-q-r}\right)^{2(k-q-r)}\\
&\prec \quad \frac{k^7n^2}{N}\lambda_1^{2k-2}\lambda_0 e^{4k-2q} \; .
\end{aligned}$$

In the third line, we bound the probability by counting the number of edges involved in the event and the number of possible configurations, as we did before. In the fourth line, we use the bound of $k/(k-q)$ to obtain a ratio of the form $\frac{k-q}{k-q-r}$. In the last line, we observe that the sum is decreasing with respect to r and is maximized at $r=0$.

Acknowledgements

We would like to thank Jacques Verstraete and Raphael Yuster for helpful discussions and references on counting k-cycles. The research of N. Verzelen is partly supported by the french Agence

Nationale de la Recherche (ANR 2011 BS01 010 01 projet Calibration). The research of E. Arias-Castro is partially supported by a grant from the Office of Naval Research (N00014-13-1-0257).

References

Aigner, M. and G. M. Ziegler (2010). *Proofs from The Book* (Fourth ed.). Berlin: Springer-Verlag.

Alon, N. (1998). Spectral techniques in graph algorithms (invited paper). In *LATIN'98: theoretical informatics (Campinas, 1998)*, Volume 1380 of *Lecture Notes in Comput. Sci.*, pp. 206–215. Berlin: Springer.

Alon, N. and S. Gutner (2010). Balanced families of perfect hash functions and their applications. *ACM Trans. Algorithms 6*(3), Art. 54, 12.

Alon, N., M. Krivelevich, and B. Sudakov (1998). Finding a large hidden clique in a random graph. In *Proceedings of the Eighth International Conference "Random Structures and Algorithms" (Poznan, 1997)*, Volume 13, pp. 457–466.

Alon, N., R. Yuster, and U. Zwick (1997). Finding and counting given length cycles. *Algorithmica 17*(3), 209–223.

Arias-Castro, E. and N. Verzelen (2012). Community detection in dense random networks. Available online at http://arxiv.org/abs/1302.7099.

Arsić, B., D. Cvetković, S. K. Simić, and M. Škarić (2012). Graph spectral techniques in computer sciences. *Appl. Anal. Discrete Math. 6*(1), 1–30.

Berthet, Q. and P. Rigollet (2012). Optimal detection of sparse principal components in high dimension. Available online at http://arXiv.org/abs/1202.5070.

Bickel, P. J. and A. Chen (2009). A nonparametric view of network models and newmanäìgirvan and other modularities. *Proceedings of the National Academy of Sciences 106*(50), 21068–21073.

Bollobás, B. (2001). *Random graphs* (Second ed.), Volume 73 of *Cambridge Studies in Advanced Mathematics*. Cambridge: Cambridge University Press.

Boucheron, S., O. Bousquet, G. Lugosi, and P. Massart (2005). Moment inequalities for functions of independent random variables. *Ann. Probab. 33*(2), 514–560.

Butucea, C. and Y. I. Ingster (2011). Detection of a sparse submatrix of a high-dimensional noisy matrix. Available from http://arxiv.org/abs/1109.0898.

Chung, F. and M. Radcliffe (2011). On the spectra of general random graphs. *Electron. J. Combin. 18*(1), Paper 215, 14.

Chung, F. R. K. (1997). *Spectral graph theory*, Volume 92 of *CBMS Regional Conference Series in Mathematics*. Published for the Conference Board of the Mathematical Sciences, Washington, DC.

d'Aspremont, A., L. El Ghaoui, M. I. Jordan, and G. R. G. Lanckriet (2007). A direct formulation for sparse pca using semidefinite programming. *SIAM Review 49*(3), 434–448.

Dekel, Y., O. Gurel-Gurevich, and Y. Peres (2011). Finding hidden cliques in linear time with high probability. In *Proceedings of the Eighth Workshop on Analytic Algorithmics and Combinatorics (ANALCO)*, pp. 67–75. Society for Industrial and Applied Mathematics (SIAM).

Feige, U. and E. Ofek (2005). Spectral techniques applied to sparse random graphs. *Random Structures Algorithms 27*(2), 251–275.

Feige, U. and D. Ron (2010). Finding hidden cliques in linear time. In *21st International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms (AofA'10)*, Discrete Math. Theor. Comput. Sci. Proc., AM, pp. 189–203. Assoc. Discrete Math. Theor. Comput. Sci., Nancy.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports 486*(3–5), 75 – 174.

Füredi, Z. and J. Komlós (1981). The eigenvalues of random symmetric matrices. *Combinatorica 1*(3), 233–241.

Girvan, M. and M. E. J. Newman (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences 99*(12), 7821–7826.

Heard, N. A., D. J. Weston, K. Platanioti, and D. J. Hand (2010). Bayesian anomaly detection methods for social networks. *Ann. Appl. Stat. 4*(2), 645–662.

Hoeffding, W. (1956). On the distribution of the number of successes in independent trials. *Ann. Math. Statist. 27*, 713–721.

Janson, S. (2005). The first eigenvalue of random graphs. *Combin. Probab. Comput. 14*(5-6), 815–828.

Krivelevich, M. and B. Sudakov (2003). The largest eigenvalue of sparse random graphs. *Combin. Probab. Comput. 12*(1), 61–72.

Lancichinetti, A. and S. Fortunato (2009, Nov). Community detection algorithms: A comparative analysis. *Phys. Rev. E 80*, 056117.

Lehmann, E. L. and J. P. Romano (2005). *Testing statistical hypotheses* (Third ed.). Springer Texts in Statistics. New York: Springer.

Maslov, S., K. Sneppen, and A. Zaliznyak (2004). Detection of topological patterns in complex networks: correlation profile of the internet. *Physica A: Statistical Mechanics and its Applications 333*(0), 529 – 540.

Mongiovı, M., P. Bogdanov, R. Ranca, E. E. Papalexakis, C. Faloutsos, and A. K. Singh (2013). Netspot: Spotting significant anomalous regions on dynamic networks. In *SIAM International Conference on Data Mining*.

Mossel, E., J. Neeman, and A. Sly (2012). Stochastic block models and reconstruction. Available from http://http://arxiv.org/abs/1202.1499.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences 103*(23), 8577–8582.

Newman, M. E. J. and M. Girvan (2004, Feb). Finding and evaluating community structure in networks. *Phys. Rev. E 69*, 026113.

Park, Y., C. Priebe, and A. Youssef (2013). Anomaly detection in time series of graphs using fusion of graph invariants. *Selected Topics in Signal Processing, IEEE Journal of 7*(1), 67–75.

Pittel, B. and N. C. Wormald (2005). Counting connected graphs inside-out. *Journal of Combinatorial Theory, Series B 93*(2), 127 – 172.

Pothen, A., H. D. Simon, and K.-P. Liou (1990). Partitioning sparse matrices with eigenvectors of graphs. *SIAM J. Matrix Anal. Appl. 11*(3), 430–452. Sparse matrices (Gleneden Beach, OR, 1989).

Reichardt, J. and S. Bornholdt (2006, Jul). Statistical mechanics of community detection. *Phys. Rev. E 74*, 016110.

Robinson, R. W. and N. C. Wormald (1992). Almost all cubic graphs are Hamiltonian. *Random Structures Algorithms 3*(2), 117–125.

Robinson, R. W. and N. C. Wormald (1994). Almost all regular graphs are Hamiltonian. *Random Structures Algorithms 5*(2), 363–374.

Rukhin, A. and C. E. Priebe (2012). On the limiting distribution of a graph scan statistic. *Communications in Statistics-Theory and Methods 41*(7), 1151–1170.

Sun, X. and A. B. Nobel (2008). On the size and recovery of submatrices of ones in a random binary matrix. *J. Mach. Learn. Res. 9*, 2431–2453.

Takács, L. (1988). On the limit distribution of the number of cycles in a random graph. *J. Appl. Probab.* (Special Vol. 25A), 359–376. A celebration of applied probability.

Van der Hofstad, R. (2012). *Random graphs and complex networks.* Available online at http:

//www.win.tue.nl/~rhofstad/NotesRGCN.pdf.

Wang, B., J. M. Phillips, R. Schreiber, D. M. Wilkinson, N. Mishra, and R. Tarjan (2008). Spatial scan statistics for graph clustering. In *SIAM International Conference on Data Mining*.

Wormald, N. C. (1999). Models of random regular graphs. In *Surveys in combinatorics, 1999 (Canterbury)*, Volume 267 of *London Math. Soc. Lecture Note Ser.*, pp. 239–298. Cambridge: Cambridge Univ. Press.